IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Po Fai Chu >
<Mar.2024>

Outline

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

Executive Summary: Overview of SpaceX Falcon 9 First Stage Landing Prediction

- Summary of methodologies

- **Data Analysis:** understand the underlying patterns and factors influencing the success of first-stage landings.
- **Visualization:** illustrate the failure rates and other critical metrics, providing clear insights into the dataset's characteristics.
- **Machine Learning Techniques:** predicting the successful landing of the Falcon 9 first stage.
- **Model Training:** enhance predictive accuracy.
- **Model Comparison:** determine the most effective approach for this specific dataset.
- **Parameter Analysis:** Identified key parameters that significantly influence the success rate.

- Summary of all results

  - **Trends:** Success rates are going up every year.
  - **Payload Insights:**
    - 6K-8K payload range has the lowest success rates.
    - 2K-4K payload range has the highest success rates.
  - **Best Model:** Decision Tree model works best for our data, found through Grid Search.
  - **Launch Sites:**
    - Most successful launches: KSC LC-39A.
    - Highest success rate: CCAFS SLC-40.
  - **Booster Version:** F9 booster version FT shows the highest success rate.

Introduction

- Project background and context
  - **SpaceX's Competitive Pricing**: Falcon 9 launches are offered at $62 million, significantly lower than competitors' costs, which can exceed $165 million.
  - **Innovation in Reusability**: The key to SpaceX's cost advantage is the reusability of the Falcon 9 rocket's first stage.
  - **Impact on Launch Costs**: Successful landing of the first stage is critical for keeping launch costs low.
- Problems to be answered
  - **Predicting Landing Success**: How can we accurately predict the successful landing of the Falcon 9 first stage?
  - **Strategic Advantage**: How does the ability to predict first stage landings impact the competitive positioning of SpaceX in bidding for launch contracts?
  - **Cost-Benefit Analysis**: What are the financial implications of successful VS unsuccessful landings on the overall cost of launches?

Section 1

Methodology

Methodology

Executive Summary

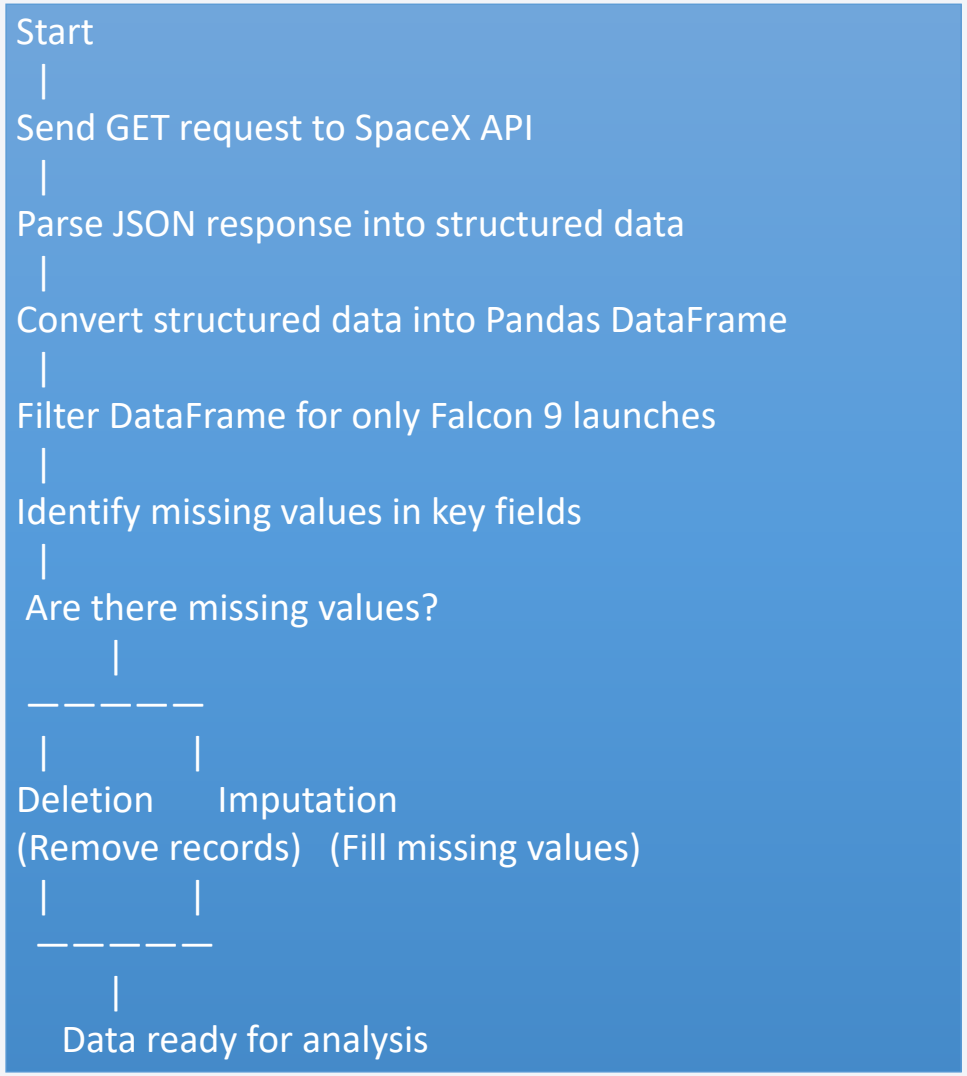
- Data collection methodology:
  - Data was meticulously gathered from SpaceX's official launch records and other reliable databases, such as scraping from wikipedia.
  - Emphasis was placed on acquiring comprehensive launch metrics, including payload, orbit types, launch sites, and outcomes.
- Perform data wrangling
  - Undertook rigorous data cleaning to correct inconsistencies, handle missing values, and remove irrelevant entries.
  - Structured data into a coherent format suitable for in-depth analysis.
  - Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built models using algorithms such as Logistic Regression, Decision Trees.
  - Tuned models through parameters optimization to enhance predictive performance.
  - Evaluated models using metrics like accuracy, precision, and recall, selecting the best performer for deployment.

Data Collection

- Utilizing SpaceX's API:
- API Access: Leveraged SpaceX's official API to directly access comprehensive data on Falcon 9 launches, including launch dates, payloads, outcomes, and more.
- Data Retrieval: Employed the Python requests library to efficiently send GET requests to the API, ensuring real-time access to the latest launch data.
- Web Scraping for Supplementary Data:
- Enhancing Data Depth: To enrich the dataset, additional information was meticulously scraped from relevant Wikipedia pages, focusing on details not covered by the API.
- Web Scraping Tools: Utilized the BeautifulSoup library in Python for web scraping, enabling precise extraction of structured data from HTML content on Wikipedia.
- Data Integration:
- Comprehensive Dataset: Combined data retrieved from SpaceX's API with information scraped from Wikipedia to create a holistic dataset for analysis.
- Data Processing: Ensured the integrity and coherence of the combined dataset through careful data cleaning and preprocessing steps.API Access: Leveraged SpaceX's official API to directly access comprehensive data on Falcon 9 launches, including launch dates, payloads, outcomes, and more.
- Data Retrieval: Employed the Python requests library to efficiently send GET requests to the API, ensuring real-time access to the latest launch data.

Data Collection – SpaceX API

- **data collection with SpaceX REST : flowcharts** 👉

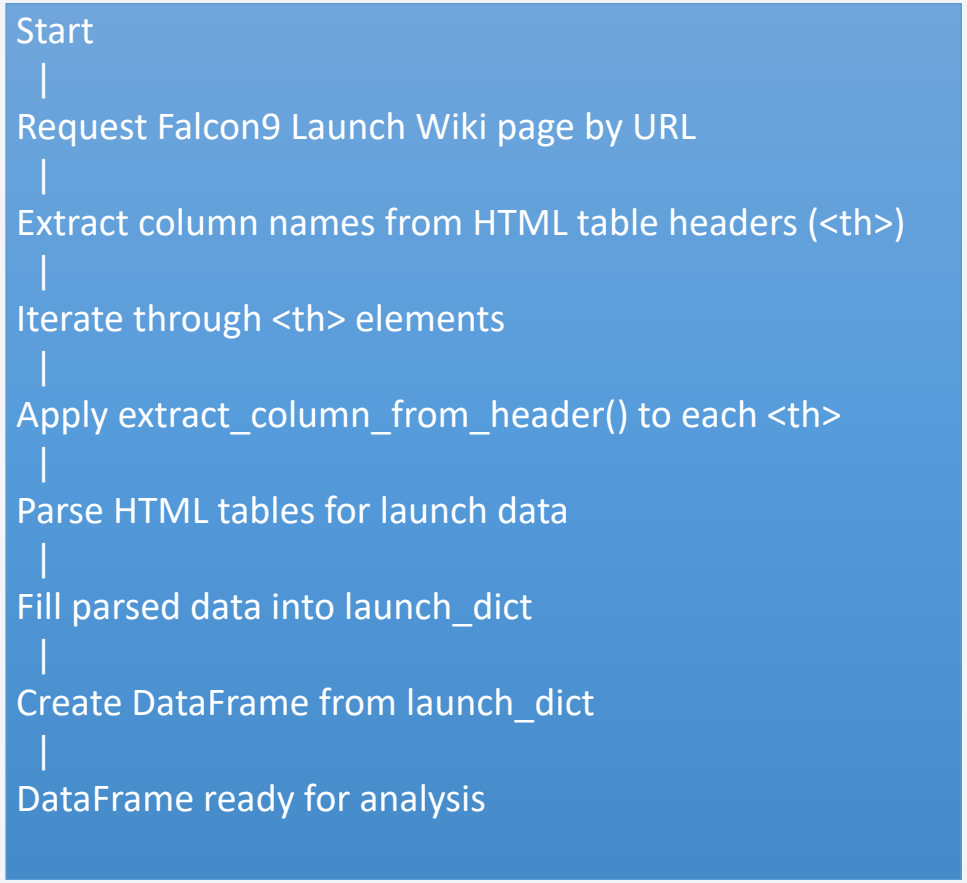


- completed SpaceX API calls notebook 🔗

Data Collection - Scraping

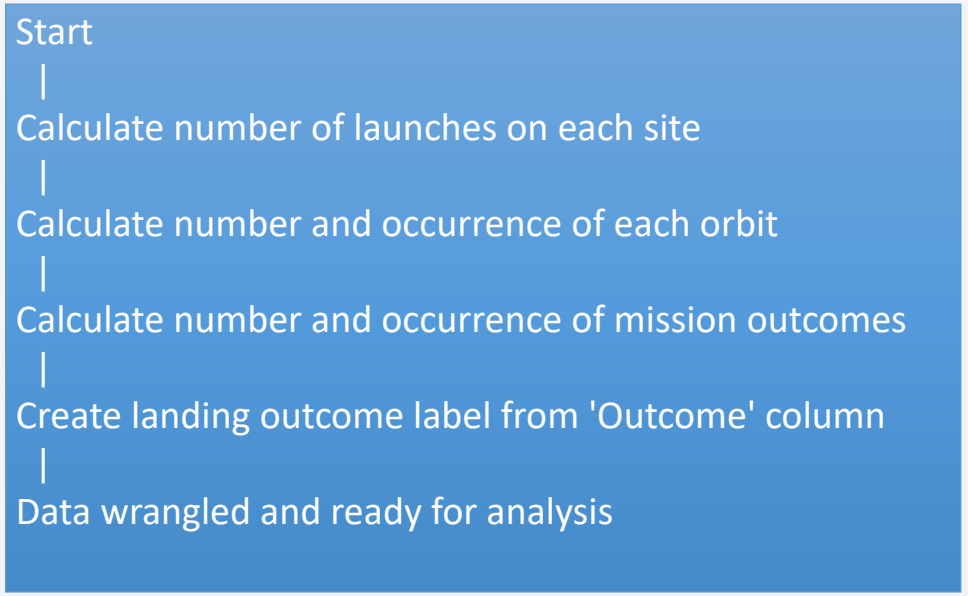
- **web scraping process flowcharts** 👉

- [GitHub URL of the completed web scraping notebook](#) 🔗



Data Wrangling

- Utilized the BeautifulSoup library in Python for web scraping, enabling precise extraction of structured data from HTML content on Wikipedia.
- **data wrangling process flowcharts** 👉



- GitHub URL of completed data wrangling 🔗

EDA with Data Visualization

- **Charts Types:**
- ***Scatter Plot:***
- Purpose: To illustrate the correlation between different variables within the Falcon 9 launch dataset.
- Why Used: Scatter plots are ideal for showing the relationship between two continuous variables, allowing you to identify patterns, trends, or any potential correlations. This makes it an excellent choice for exploring how various factors might interact with each other, such as launch payload vs. success rate.
- ***Bar Chart:***
- Purpose: To depict the success rates associated with each orbit type.
- Why Used: Bar charts are highly effective for comparing numerical values across different categories. In this case, using a bar chart enables a clear visualization of how success rates vary by orbit, highlighting which orbits have higher or lower success rates. This visualization is straightforward for audiences to understand and provides immediate insights into orbit-specific performance.
- ***Line Chart:***
- Purpose: To track the annual success rates of Falcon 9 launches.
- Why Used: Line charts are particularly suited for displaying data trends over time. Employing a line chart to monitor annual success rates allows for an easy assessment of whether the success rate is improving, declining, or remaining stable over the years. This can be critical for evaluating the effectiveness of improvements made to the launch process or technology over time.
- GitHub URL of your completed EDA with data visualization 🔗

EDA with SQL

- **Retrieve Total Payload Mass for NASA CRS**:Query to sum up the total payload mass carried by all missions under NASA's Commercial Resupply Services (CRS).
- A**verage Payload Mass of Booster Version F9 v1.1:**Query to calculate the average payload mass carried by the Falcon 9 version 1.1 booster.
- **Date of First Successful Landing:**Query to identify the earliest date on which a successful landing was achieved.
- T**otal Number of Successful and Failed Missions:**Query to count the total number of missions that were successful and those that failed, analyzing the mission outcomes.
- **Booster Versions with Maximum Payload Mass:**Query to identify the booster versions that have transported the maximum payload mass to date.

- GitHub URL of completed EDA with SQL notebook 🔗

Build an Interactive Map with Folium

- **Summary of Folium Map Enhancements**
- ***Map Creation and Site Highlighting:***
- Initial Map: Centered at NASA Johnson Space Center, Houston, Texas.
- Launch Site Circles: Highlighted each launch site with folium.Circle based on coordinates, visually emphasizing their locations.
- ***Launch Outcome Visualization:***
- Markers for Launch Records: Indicated success or failure with color-coded markers, differentiating outcomes for easy visual analysis.
- ***Proximity Analysis:***
- Distance Measurements: Utilized MousePosition to identify points of interest, calculating distances to coastlines, cities, railways, and highways.
- Visualization: Marked points and drew lines to show distances, providing insights into the strategic positioning of launch sites.

- **Rationale for Map Objects**
- Geospatial Context: Markers and circles offer a clear understanding of SpaceX launch sites and outcomes.
- Strategic Insights: Analyzing distances between sites and proximities reveals logistical and strategic site selection considerations.
- Interactive Exploration: Interactive elements like distance markers and lines foster user engagement, allowing for a deeper exploration of spatial relationships.

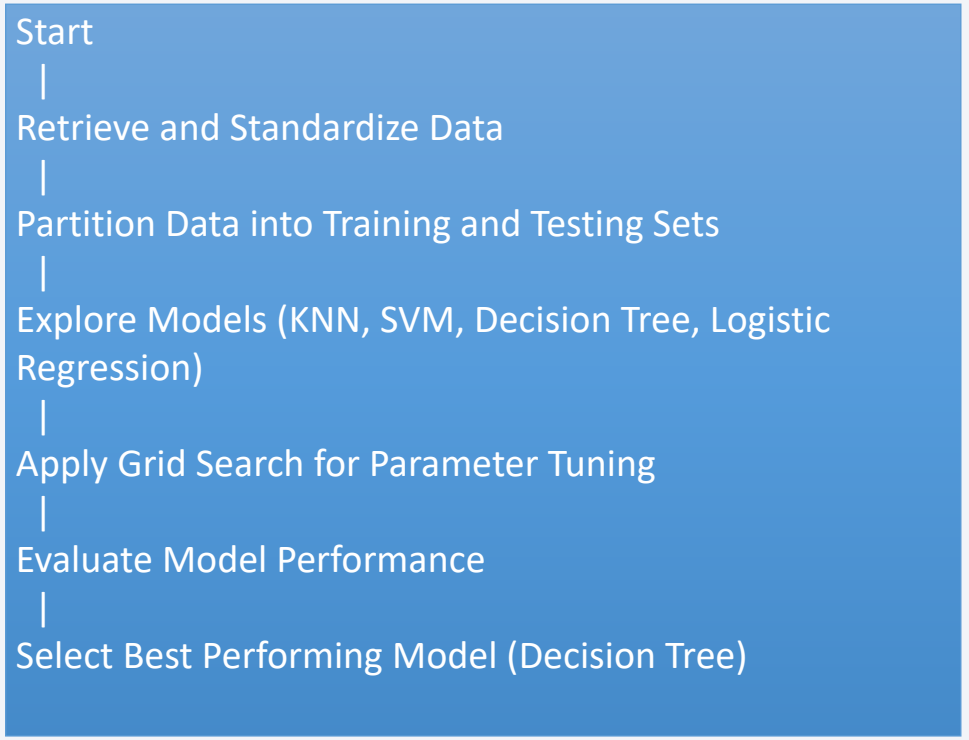
- GitHub URL of completed interactive map with Folium map 🔗

Build a Dashboard with Plotly Dash

- **Dashboard Enhancements Summary**
- **Plots/Graphs and Interactions Added:**
- **1.Launch Site Selection Dropdown:**
- Allows users to select a launch site or view data for all sites.
- Enhances user interaction by filtering the dataset according to the selected site.
- **2.Success Rate Pie Chart:**
- Displays the total successful launches count for all sites or for a specific site.
- Provides a visual representation of launch success rates, making it easy to compare outcomes across sites or assess a single site's performance.
- **3.Payload Range Slider:**
- Enables users to select a payload mass range for further analysis.
- Facilitates exploration of launch success correlation with different payload masses.
- **4.Payload vs. Launch Success Scatter Chart:**
- Illustrates the correlation between payload mass and launch success for all sites or a selected site.
- Offers insights into how payload mass impacts launch outcomes, highlighting potential trends or patterns.
- **Reasons**
- **Enhanced User Engagement:** The dropdown and slider allow users to interact with the dashboard actively, tailoring the displayed information to their interests.
- **Insightful Visualizations:** The pie and scatter charts provide immediate visual feedback on the success rates and potential correlations, enabling users to quickly grasp complex data relationships.
- **Data-Driven Decisions**: By visualizing success rates and examining the payload's impact on success, stakeholders can make informed decisions regarding future launches.
- **Comparative Analysis:** The ability to filter by launch site and payload range offers a comparative view, aiding in identifying which factors contribute most significantly to mission success.
- GitHub URL of completed Plotly Dash lab 🔗

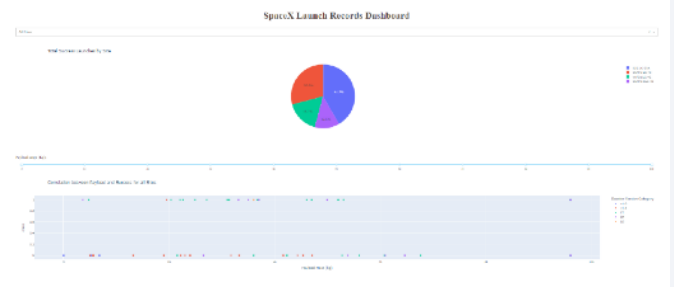
Predictive Analysis (Classification)

- **Model Development Summary**
- **Process Overview:**
- 1.Data Preparation: Standardized the dataset for uniformity and split it into training and testing sets for evaluation.
- 2.Model Exploration: Assessed various models (KNN, SVM, Decision Tree, Logistic Regression) using Grid Search for optimal parameter tuning.
- 3.Optimal Model Selection: Identified the Decision Tree model as the best performer due to its superior accuracy.
- **Key Steps for Optimization**:
- Utilized Grid Search to fine-tune model parameters, enhancing performance.
- Selected the Decision Tree model based on rigorous accuracy evaluation.
- **Process Flowchart** 👉
- GitHub URL of completed predictive analysis lab 🔗



Results

- **Exploratory data analysis results**
- Key Findings from Dashboard Analytics:
- Payload Range Insights: The 6K-8K kg payload range exhibits the lowest success rates, indicating potential challenges or limitations within this weight class.
- Model Optimization: Utilizing Grid Search enabled the identification of optimal model parameters, revealing the Decision Tree model as the most accurate for predicting launch outcomes.
- Launch Site Performance:
- KSC LC-39A stands out with the largest number of successful launches, showcasing its reliability.
- CCAFS SLC-40 boasts the highest launch success rate, highlighting its efficiency.
- Payload Success Rates: The 2K-4K kg payload range achieves the highest launch success rate, suggesting an optimal range for successful missions.
- Booster Version Efficiency: The Falcon 9 (F9) booster version FT demonstrates the highest launch success rate, indicating superior performance.
- **Predictive analysis results**
- Model Assessment: Evaluated multiple classification models (KNN, SVM, Decision Tree, Logistic Regression) to predict outcomes based on the dataset.
- Optimal Model Selection: The Decision Tree model was identified as the most effective, outperforming others in terms of accuracy through Grid Search optimization.
- Accuracy Achievement: The selected Decision Tree model demonstrated superior predictive performance, making it the preferred choice for this analysis.

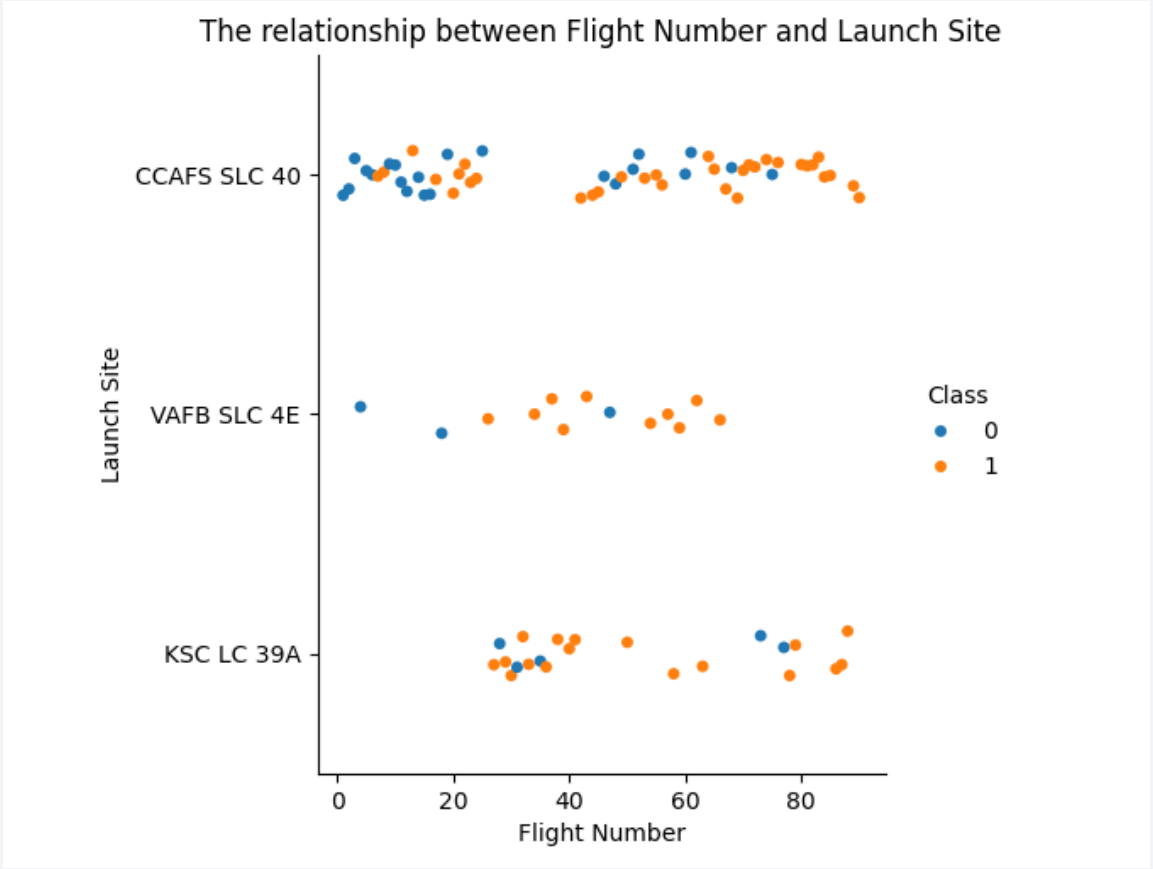


Section 2

Insights drawn from EDA

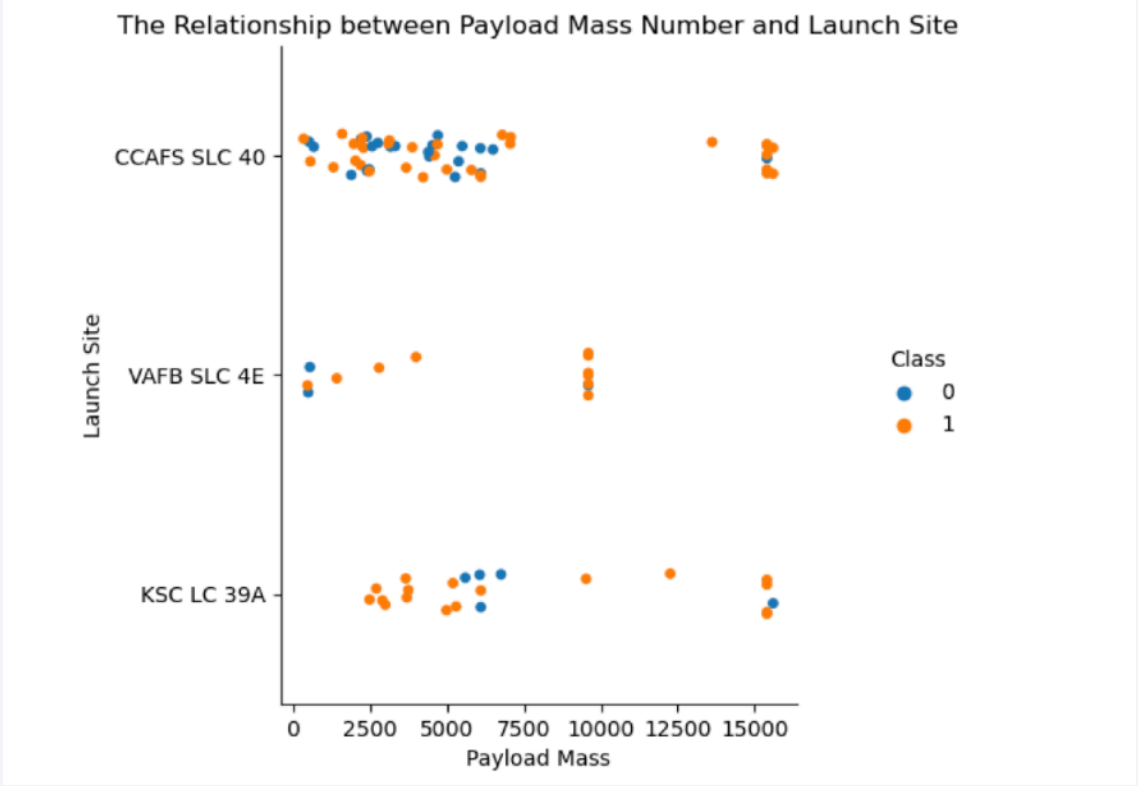
Flight Number vs. Launch Site

- the Flight Number of VAFB-SLC 4E is less than others.
- The CCAFS SLC held the most flights



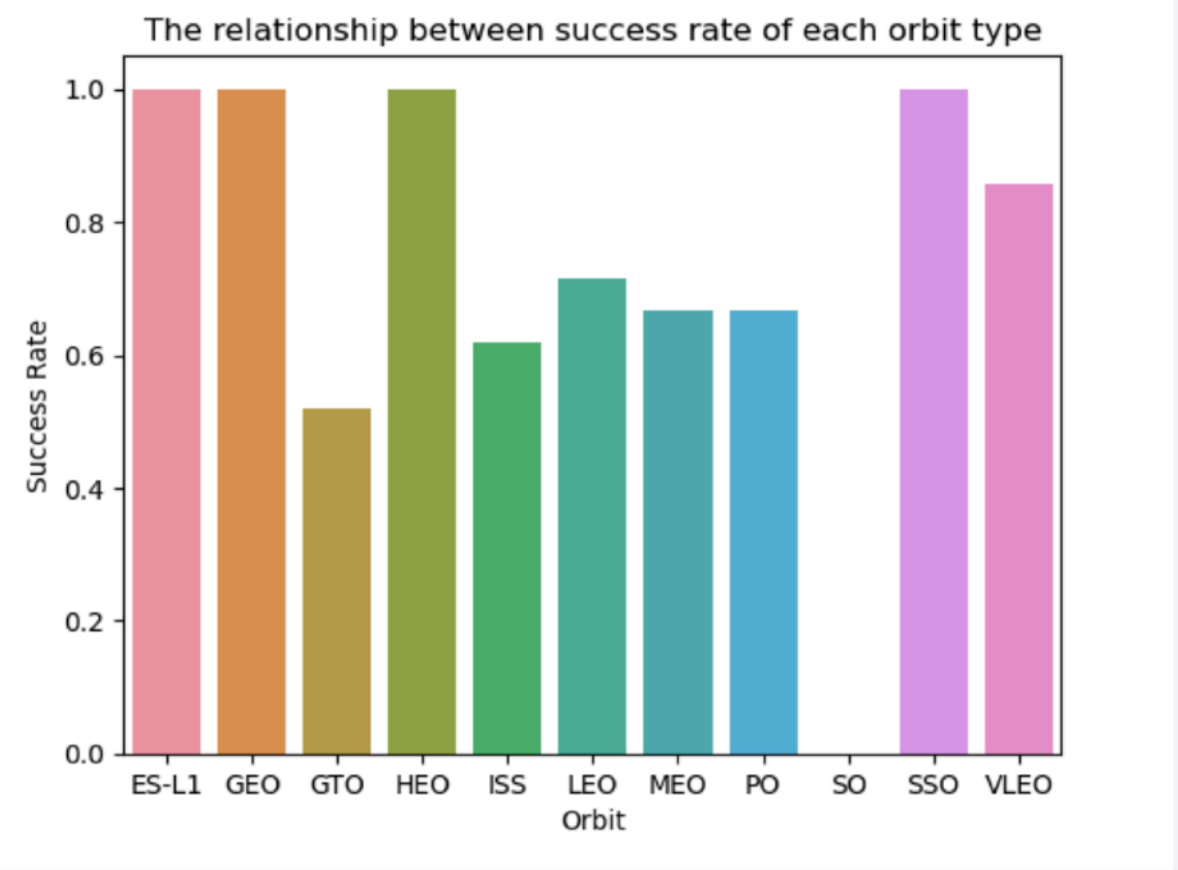
Payload vs. Launch Site

- the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).



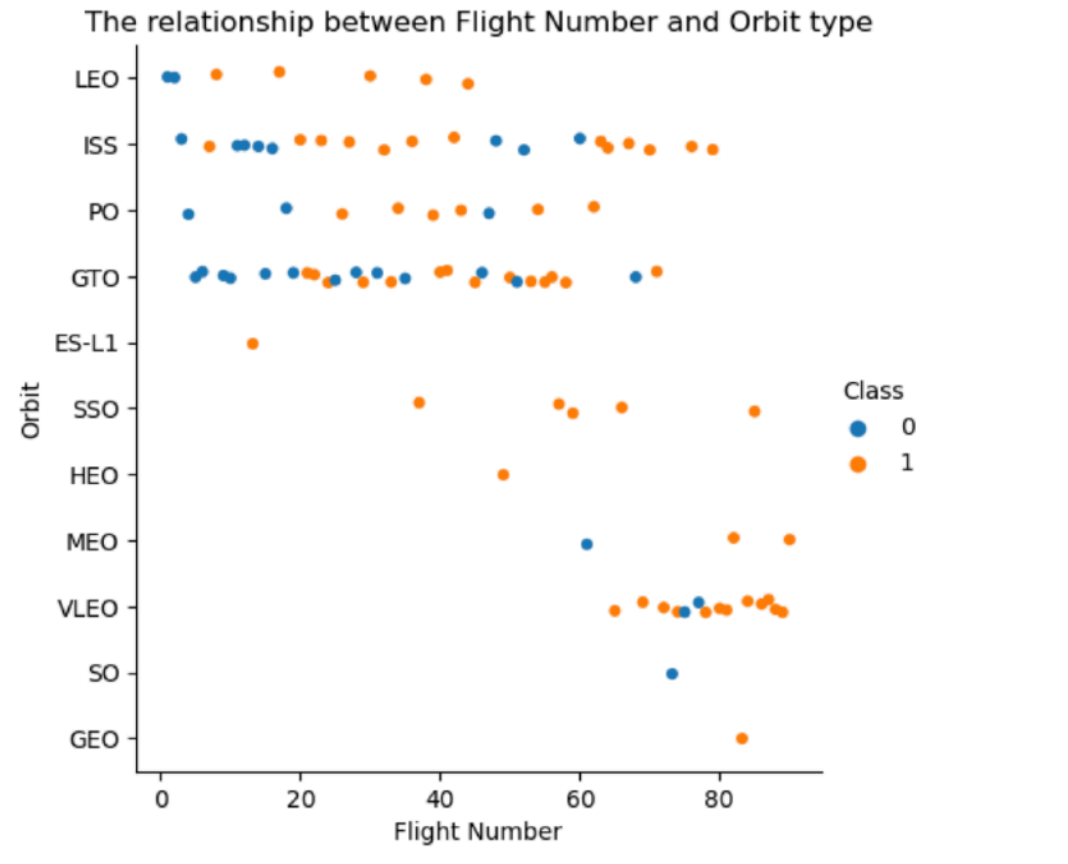
Success Rate vs. Orbit Type

- orbits have high success rate:
- ES-L1,GEO,HEO,SSO
- Lowest success rate: SO



Flight Number vs. Orbit Type

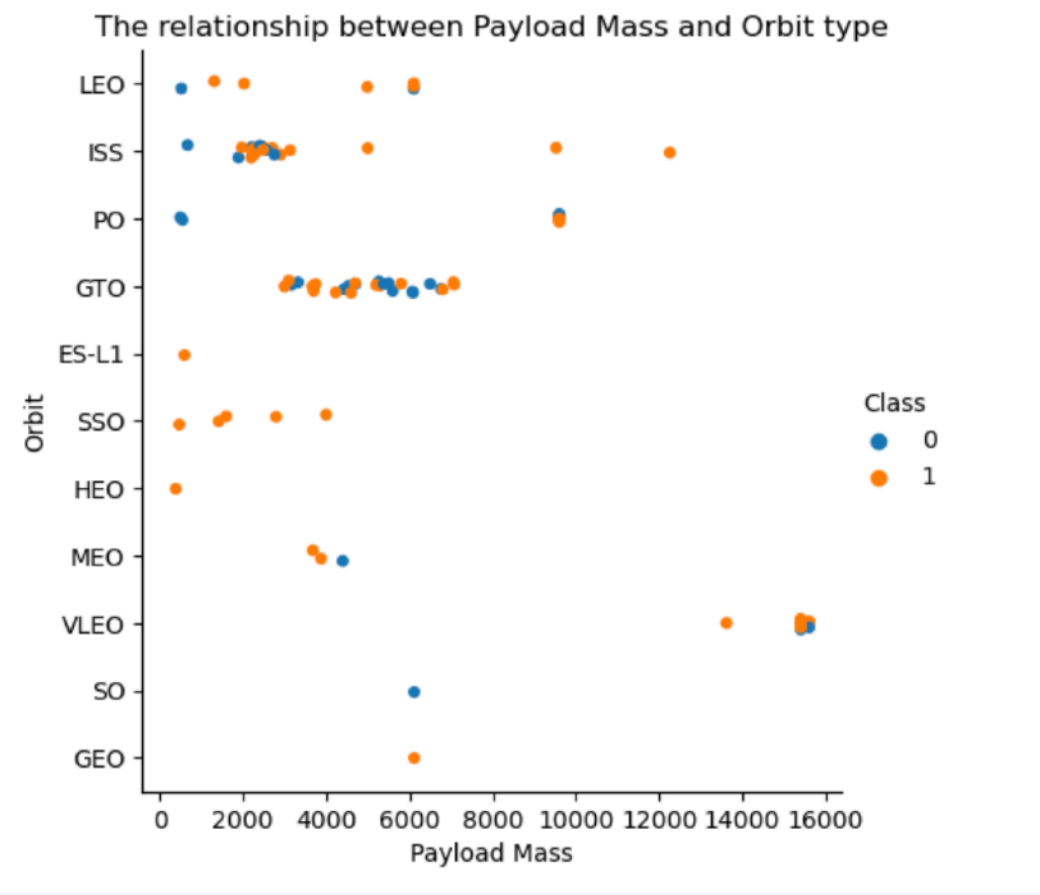
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

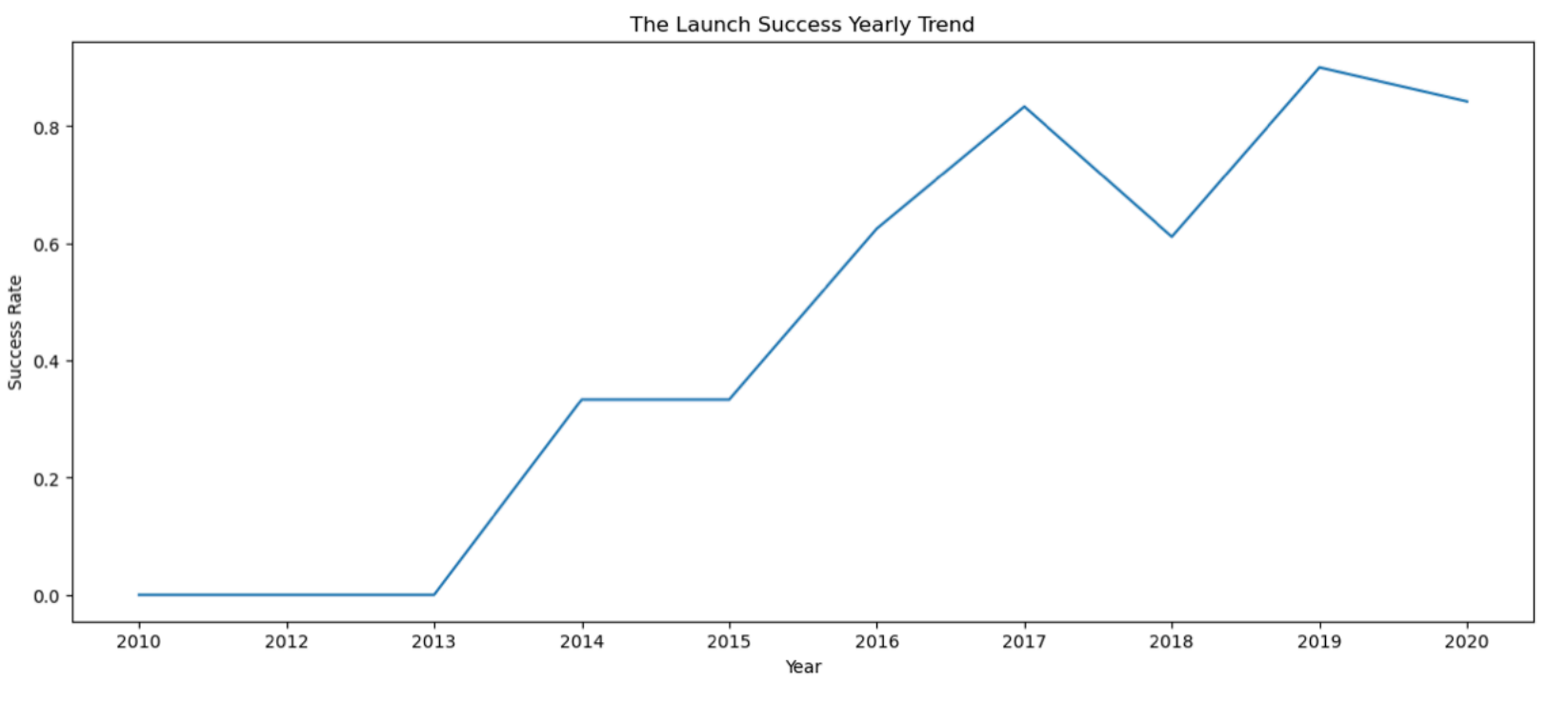
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

- the unique launch sites

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

- %sql select DISTINCT Launch_Site FROM SPACEXTABLE

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

- `%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site like 'CCA%' LIMIT 5`

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'

-

Average Payload Mass by F9 v1.1

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version like 'F9 v1.1%'

First Successful Ground Landing Date

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

min(Date)
2010-06-04

- %sql SELECT min(Date) FROM SPACEXTABLE WHERE Mission_Outcome = ‘Success'

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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 Mission_Outcome = 'Success' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

- %sql SELECT Mission_Outcome, COUNT(*) AS Mission_Outcome_Count FROM SPACEXTABLE GROUP BY Mission_Outcome

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)

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

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

- %%sql
- SELECT substr(Date, 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome = 'Failure (drone ship)'

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

Landing_Outcome	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

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) as Landing_Outcome_Count
FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Landing_Outcome_Count DESC
```

Section 3

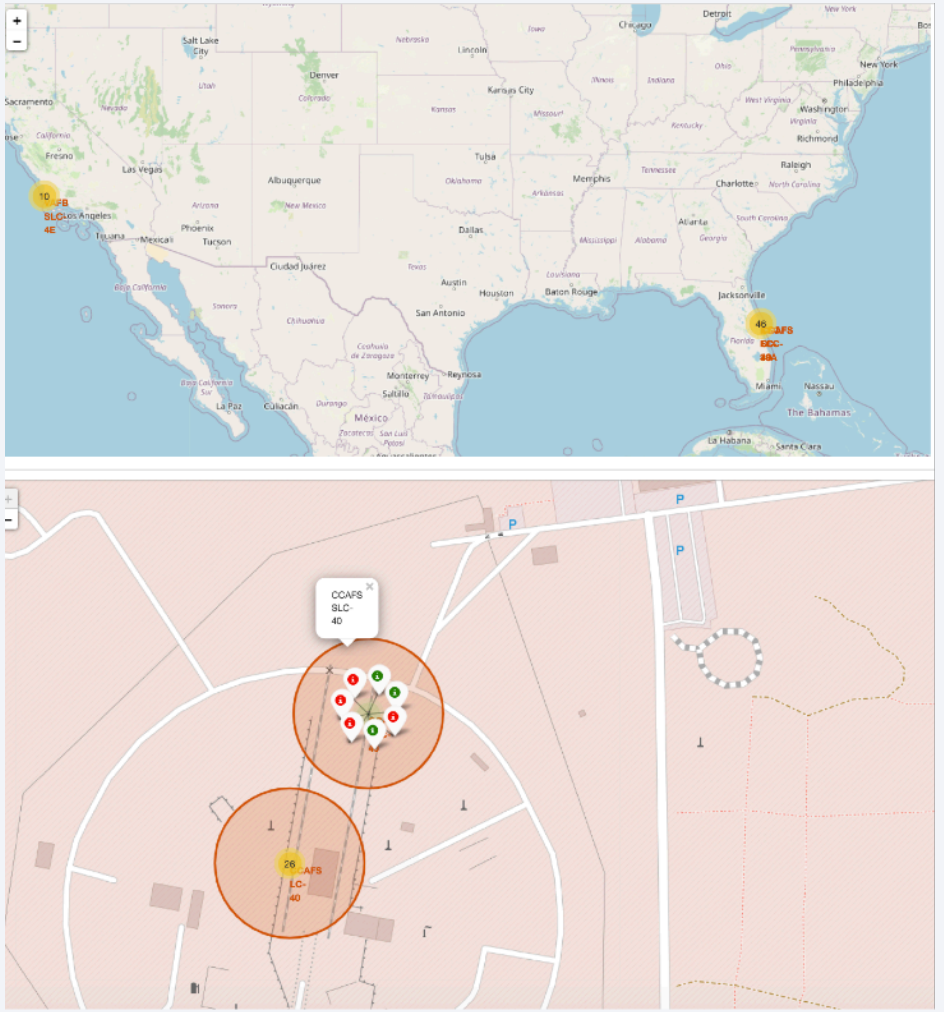
Launch Sites Proximities Analysis

<Folium Map > marked launch sites

- all launch sites in proximity to the Equator line, all launch sites in very close proximity to the coast

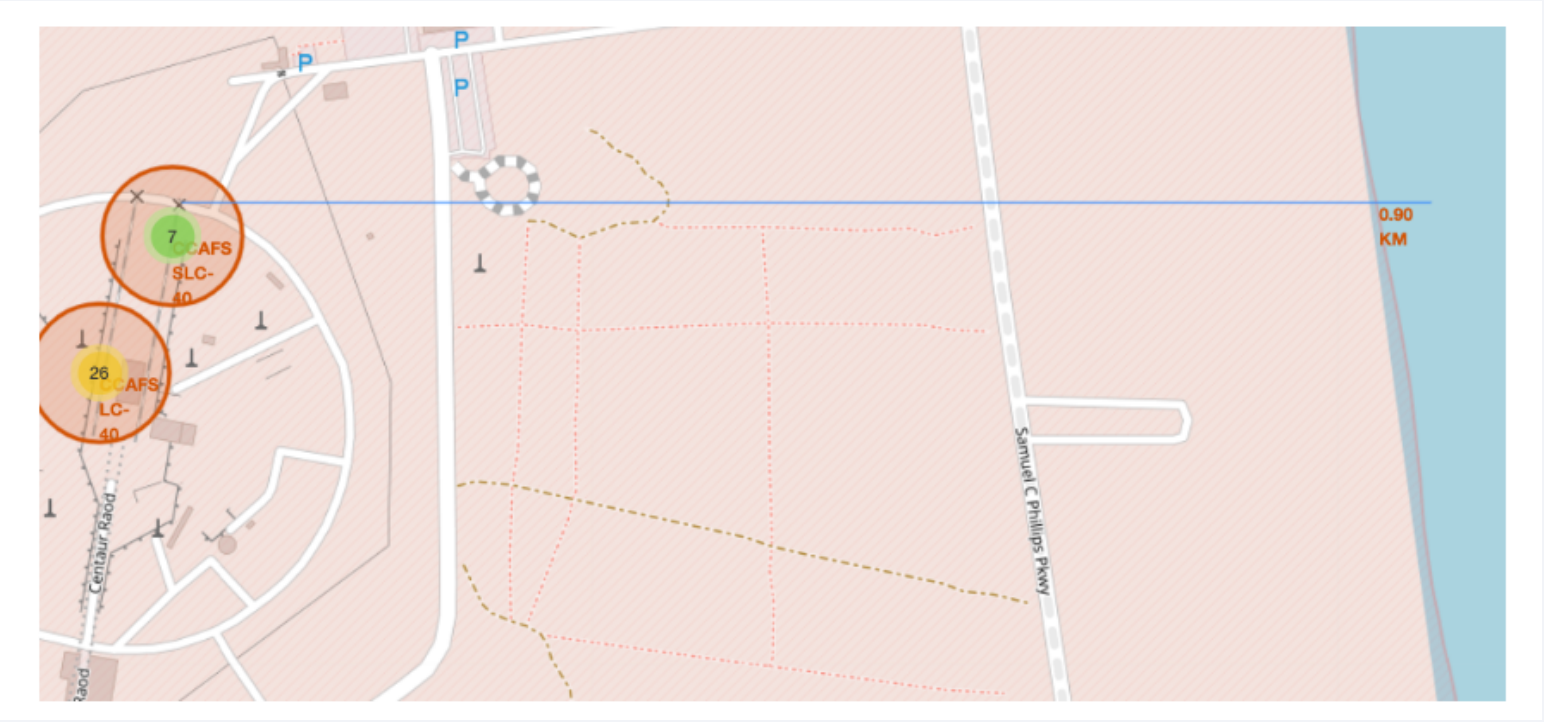
<Folium Map Launch>

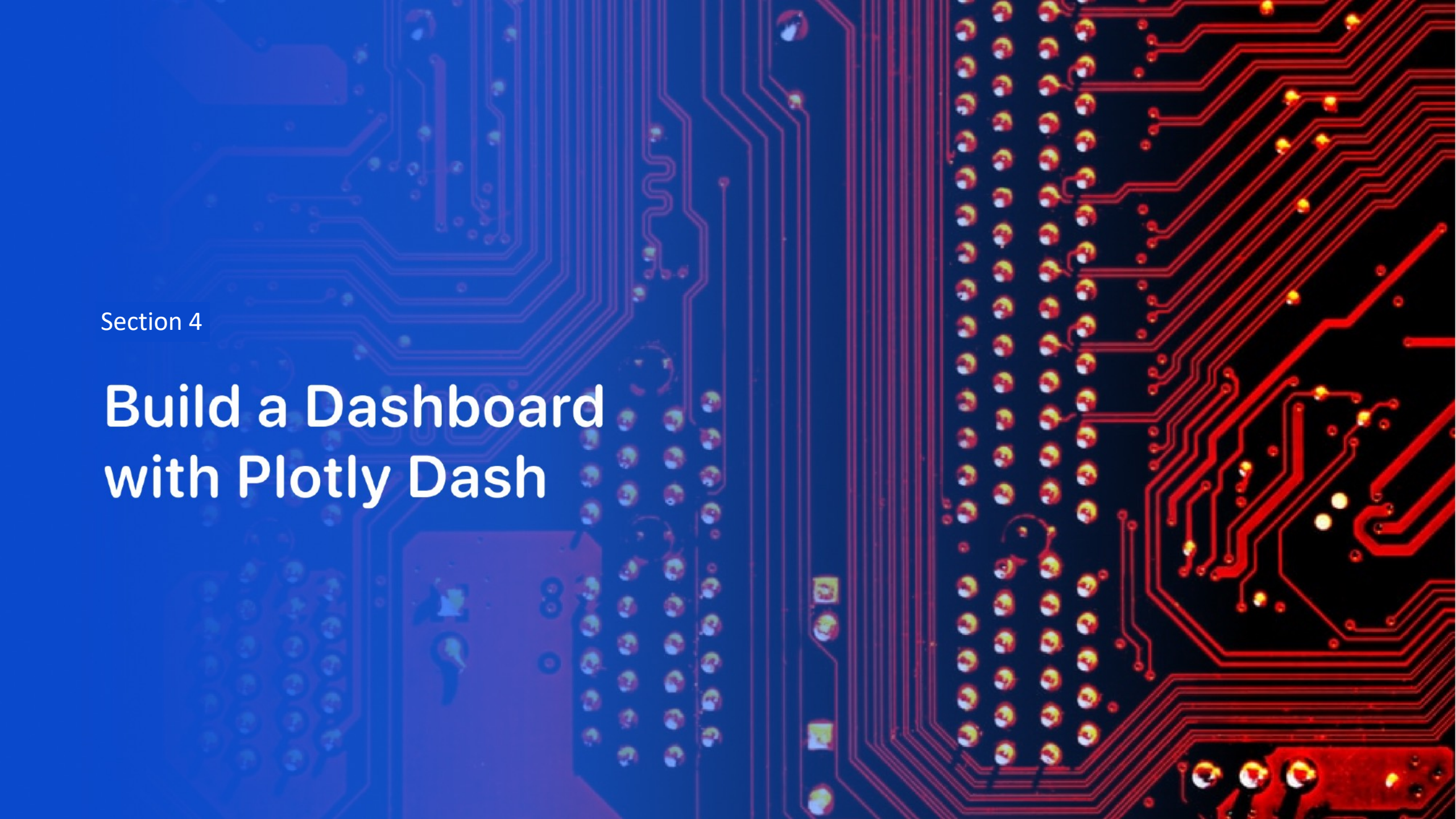
- From the color-labeled markers in marker clusters, we should be able to easily identify CCAFS SLC-40 has relatively high success rates.



<Folium Map >

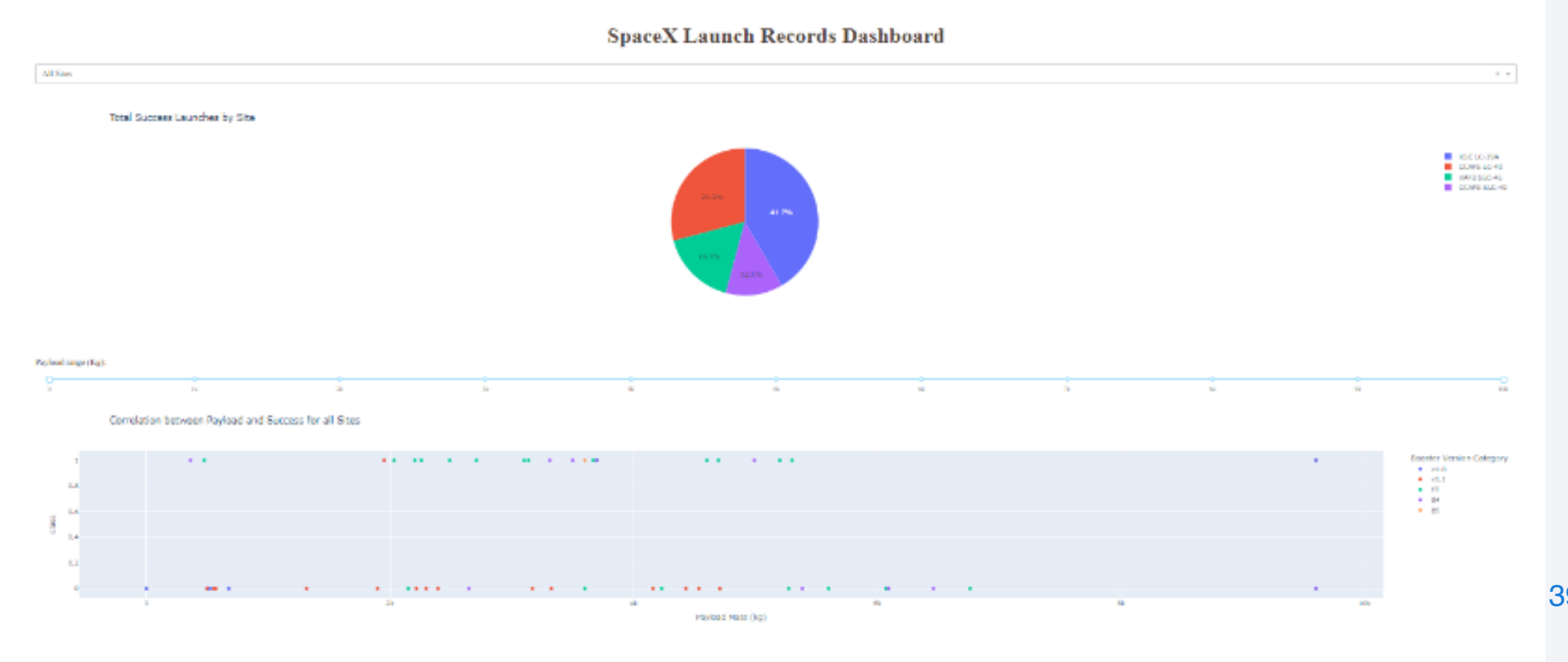
- The distance showed




Section 4
Build a Dashboard with Plotly Dash

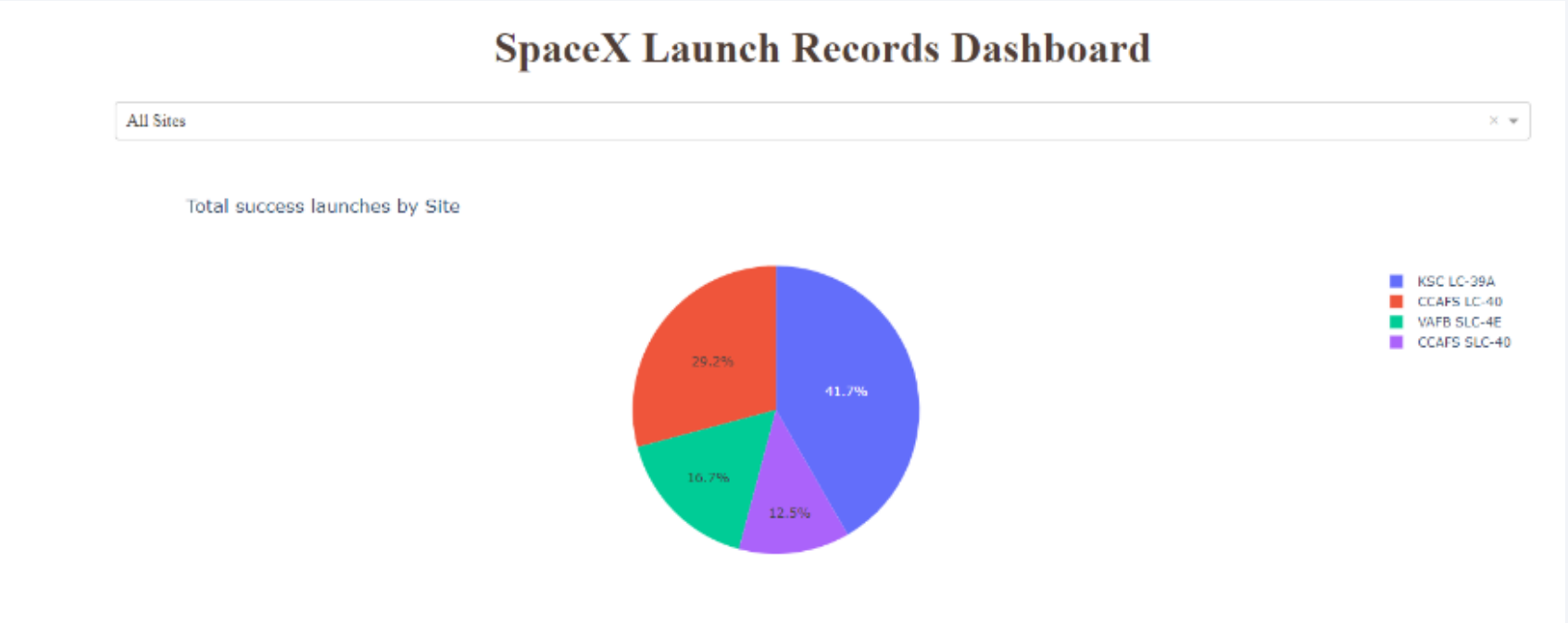
<Dashboard Overview>

- Results Overviews can be defined by the charts



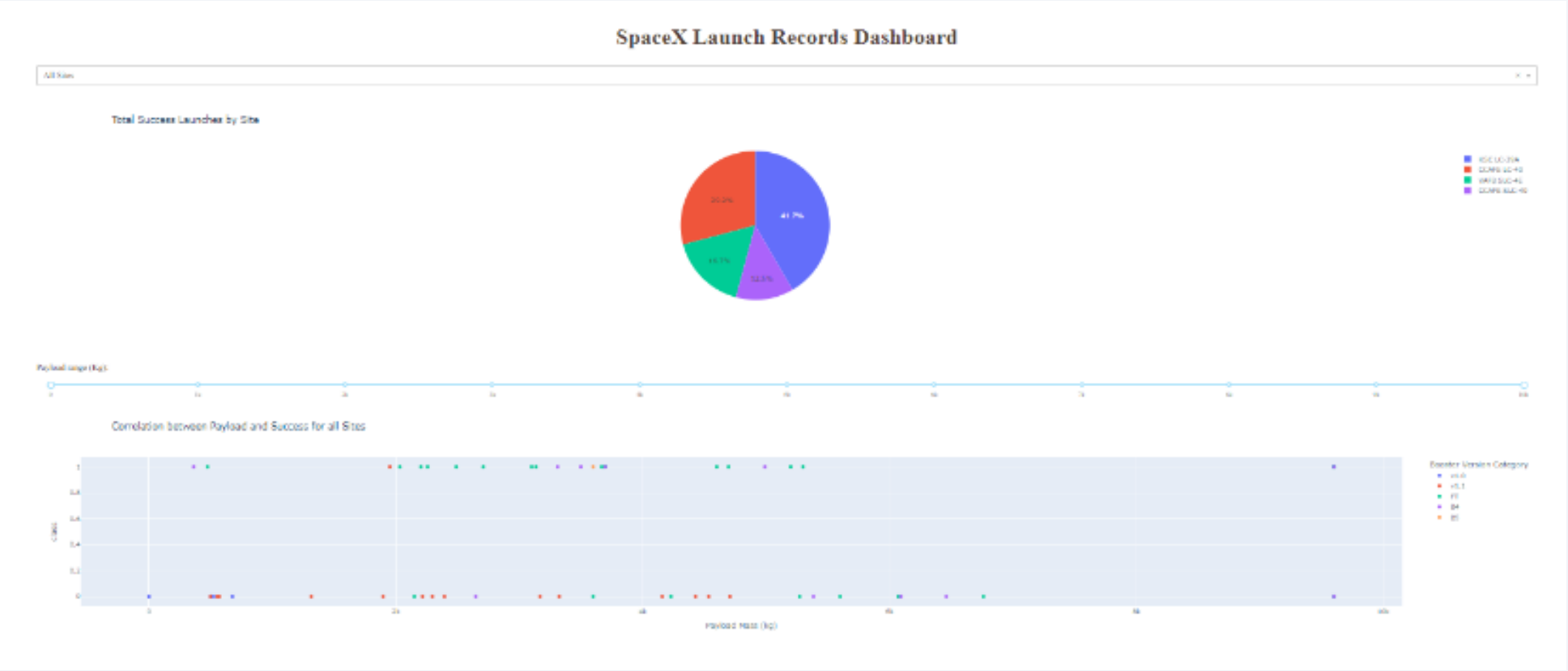
<Dashboard > SpaceX Launch Site & Success Ratio

- the piechart for the launch site with highest launch success ratio



- KSC LC took the highest launch success ratio

<Dashboard >Payload vs. Launch Outcome

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- 

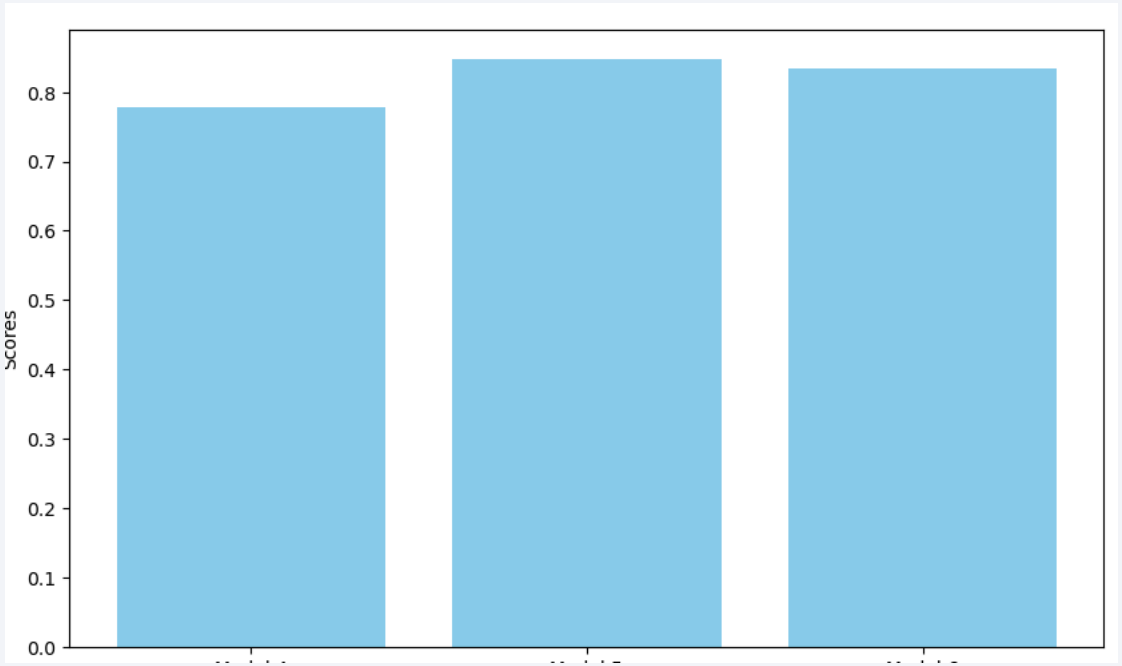
Section 5

Predictive Analysis (Classification)

Classification Accuracy

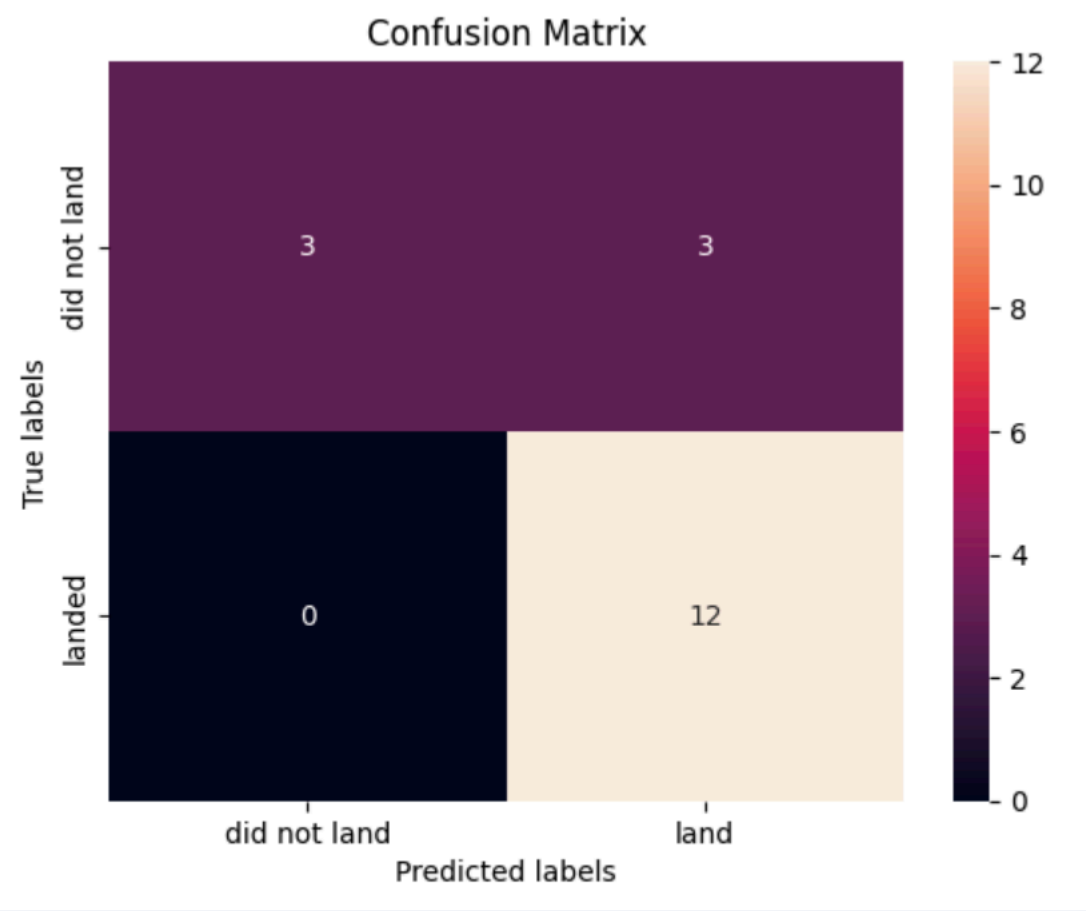
- Visualize the built model accuracy for all built classification models, in a bar chart

- k nearest neighbors has the highest classification accuracy



Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



Conclusions

- **Yearly Improvement in Success Rate:** Analysis reveals a positive trend, with SpaceX's launch success rate increasing annually, indicating ongoing improvements in launch technology and operations.
- **Payload Range Impact:** The 6K-8K kg payload range experiences the lowest success rates, suggesting challenges or limitations with heavier payloads.
- **Launch Site Performance:** Kennedy Space Center's LC-39A site leads with the highest number of successful launches, underscoring its significance in SpaceX's operations.
- **Launch Success Rates by Site:** Cape Canaveral Air Force Station's SLC-40 boasts the highest launch success rate, highlighting its operational excellence.
- **Optimal Payload Range for Success:** Launches with payloads in the 2K-4K kg range have the highest success rates, indicating an optimal payload mass for mission success.
- **Booster Version Efficiency:** The Falcon 9 booster version FT achieves the highest success rate, affirming its reliability and performance.
- **Strategic Insights for Payload Planning**: The findings suggest strategic planning regarding payload mass can significantly influence mission success, particularly highlighting the efficiency of certain payload ranges over others.
- **Evidence of Technological Advancement:** The year-over-year increase in launch success rates reflects SpaceX's commitment to technological innovation and operational improvement.
- **Guidance for Future Launch Strategy**: These insights provide valuable guidance for future SpaceX launch strategies, emphasizing the importance of selecting optimal payload ranges and booster versions to maximize success rates.

Thank you!