IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Touchthunwa Butaralambun
31 August 2023

Outline

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

Executive Summary

In this culminating project, our focus revolves around the predictive assessment of the successful landing of SpaceX Falcon 9's first stage. By establishing the capability to anticipate the outcome of the first stage landing, we unlock insights crucial to estimating launch costs. This endeavor is underpinned by the application of diverse machine learning classification algorithms.

Our methodology encompasses key stages including Data Acquisition, Data Cleansing and Preprocessing, Exploratory Data Analysis, Data Visualization, and culminates in Machine Learning Prediction. Throughout our investigative journey, compelling patterns emerge from our analysis, shedding light on specific facets of rocket launches that exhibit a discernible correlation with their success or failure.

Ultimately, our findings lead us to assert that the Decision Tree algorithm stands out as the optimal choice for tackling this intricate challenge. By leveraging its strengths, we equip ourselves with a powerful tool to navigate the complexities inherent in predicting the outcome of SpaceX Falcon 9's first stage landings.

Introduction

- At the core of this capstone project lies a pivotal objective: to anticipate the outcome of a successful landing for the Falcon 9 first stage. Notably, SpaceX has carved a niche in its ability to recurrently deploy the initial stage of their rocket launches, a distinction they proudly spotlight on their website. While conventional providers grapple with launch costs exceeding 165 million dollars, SpaceX boasts a remarkable 62 million-dollar price tag, a testament to the cost-saving potential of first stage reusability. The pivotal link between successful landings and launch expenses underscores the significance of our predictive endeavor.
- The ramifications extend beyond SpaceX's realm, carrying implications for competitive bids in the aerospace arena. As potential contenders seek to challenge SpaceX's dominance in rocket launches, the ability to gauge first stage landing outcomes becomes a strategic asset. Consequently, our central inquiry emerges: Can we leverage a curated array of Falcon 9 launch features to forecast the triumphant landing of its initial stage? This question forms the cornerstone of our exploration, propelling us to unravel the intricate dynamics underlying the success of these pioneering rocket landings.

Section 1

Methodology

Methodology

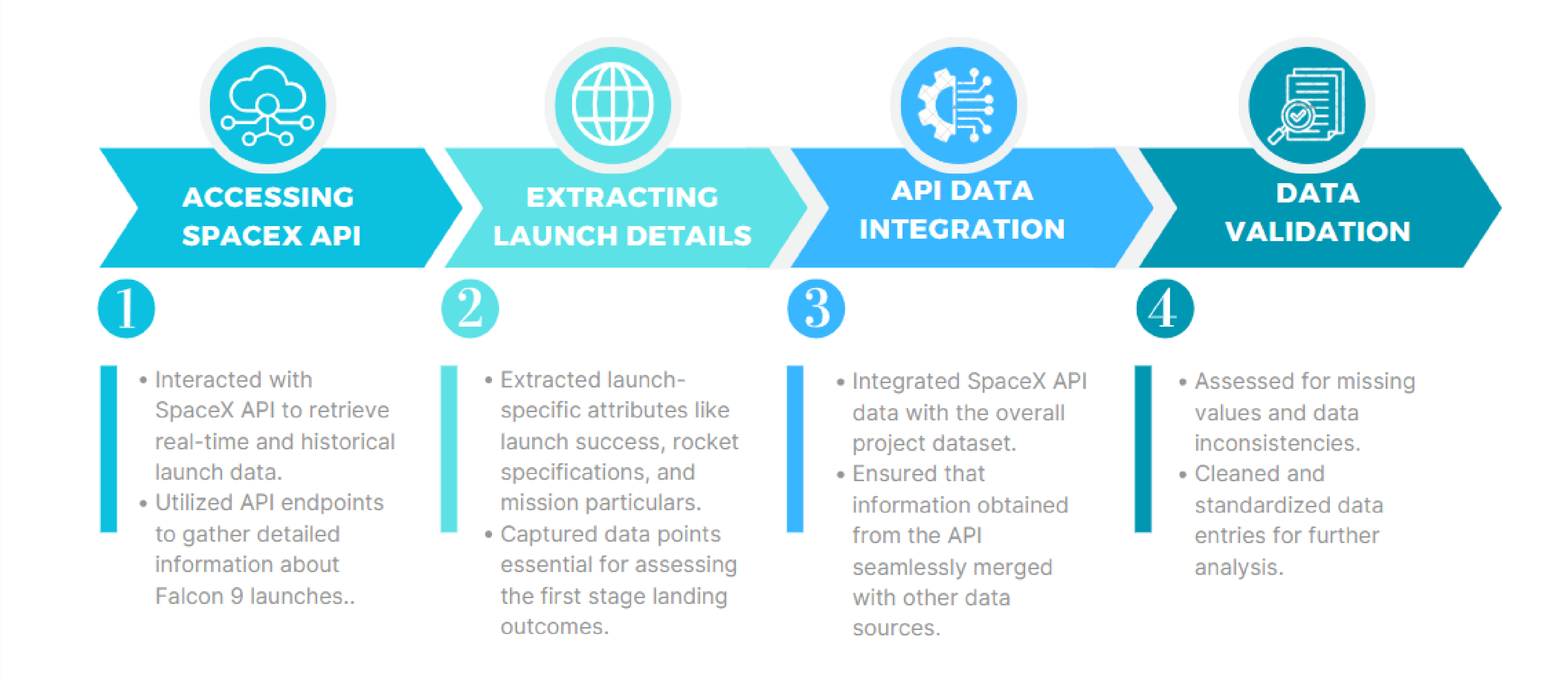
Executive Summary

- Data collection methodology
  - We employed a dual-method approach, gathering data from the SpaceX API and scraping a dedicated Wikipedia page. This ensured a comprehensive dataset.
- Perform data wrangling
  - Using Python's pandas, we refined the collected data, rectifying inconsistencies and ensuring uniformity.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We tested four classification models - logistic regression, support vector machines, k-nearest neighbors, and decision tree classifier - through systematic building, tuning, and evaluation.

Data Collection

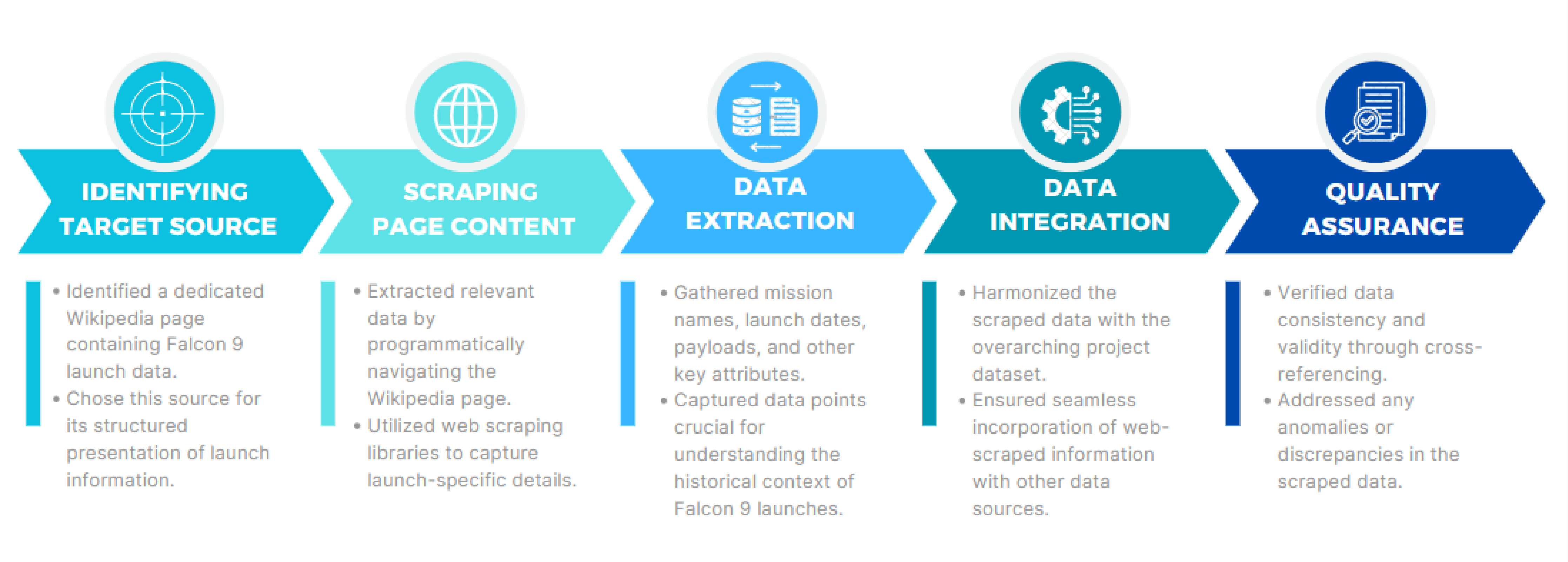
- Our data collection strategy encompassed two primary approaches.
- Firstly, pertinent data was extracted through direct interaction with the SpaceX API.
- Additionally, we augmented our dataset by scraping launch data from a dedicated Wikipedia page.
- This dual-source methodology ensured a comprehensive and diverse dataset for analysis.

Data Collection – SpaceX API



GitHub : [Data Collection]

Data Collection - Scraping



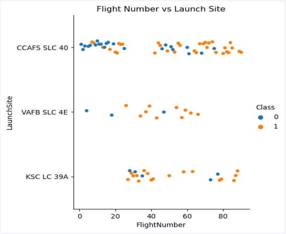
GitHub : [WebScraping]()

Data Wrangling

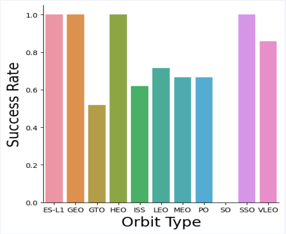
GitHub : DataWragling

EDA with Data Visualization

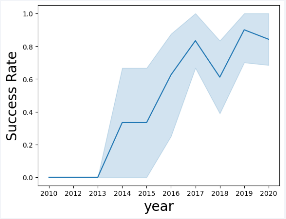
- Scatter Plots:
  - Visualized relationships between variables like Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type, and Payload vs. Orbit Type. Aided in identifying potential correlations.



- Bar Chart:
  - Enabled easy comparison of Success Rates for different Orbit Types. Conveyed variations in success across orbits efficiently.



- Line Chart:
  - Illustrated Success Rate trends over years. Offered a clear depiction of how success evolved over time.



GitHub : [EDA_DataVisualization]

EDA with SQL

- Displayed unique launch site names in space missions.
- Displayed 5 records where launch sites start with the string 'CCA'.
- Calculated the total payload mass carried by boosters launched by NASA (CRS).
- Calculated the average payload mass carried by booster version F9 v1.1.
- Listed the date of the first successful landing outcome on a ground pad.
- Listed the names of boosters with success in drone ship landings, carrying payload mass between 4000 and 6000.
- Counted the total number of successful and failed mission outcomes.
- Listed the booster versions carrying the maximum payload mass.
- Listed failed landing outcomes in drone ships, along with booster versions and launch site names in the year 2015.
- Ranked the count of landing outcomes between June 4, 2010, and March 20, 2017, in descending order.
- These queries provided insights into various aspects of Falcon 9 launch outcomes, booster versions, and their associated details.

GitHub : [EDAwithSQL]

Build an Interactive Map with Folium

- **Markers: Launch Sites**
  - Blue circle marker added at NASA Johnson Space Center's coordinate with a popup label displaying its name based on latitude and longitude.
  - Red circle markers added at all launch sites' coordinates with popup labels indicating their names using latitude and longitude.
- **Colored Markers: Launch Outcomes**
  - Added colored markers (green for successful, red for unsuccessful) at each launch site to visualize launch outcomes. This showcases which launch sites have higher success rates.
- **Colored Lines: Distances to Proximities**
  - Utilized colored lines to depict distances from launch site **CCAFS SLC-40** to its nearest coastline, railway, highway, and city. This visual representation conveys launch site proximity effectively.

These map objects and annotations collectively provide a comprehensive visualization of Falcon 9 launch sites, their outcomes, and their spatial relationships to key features.

GitHub :Interactive Map Analytics with Folium

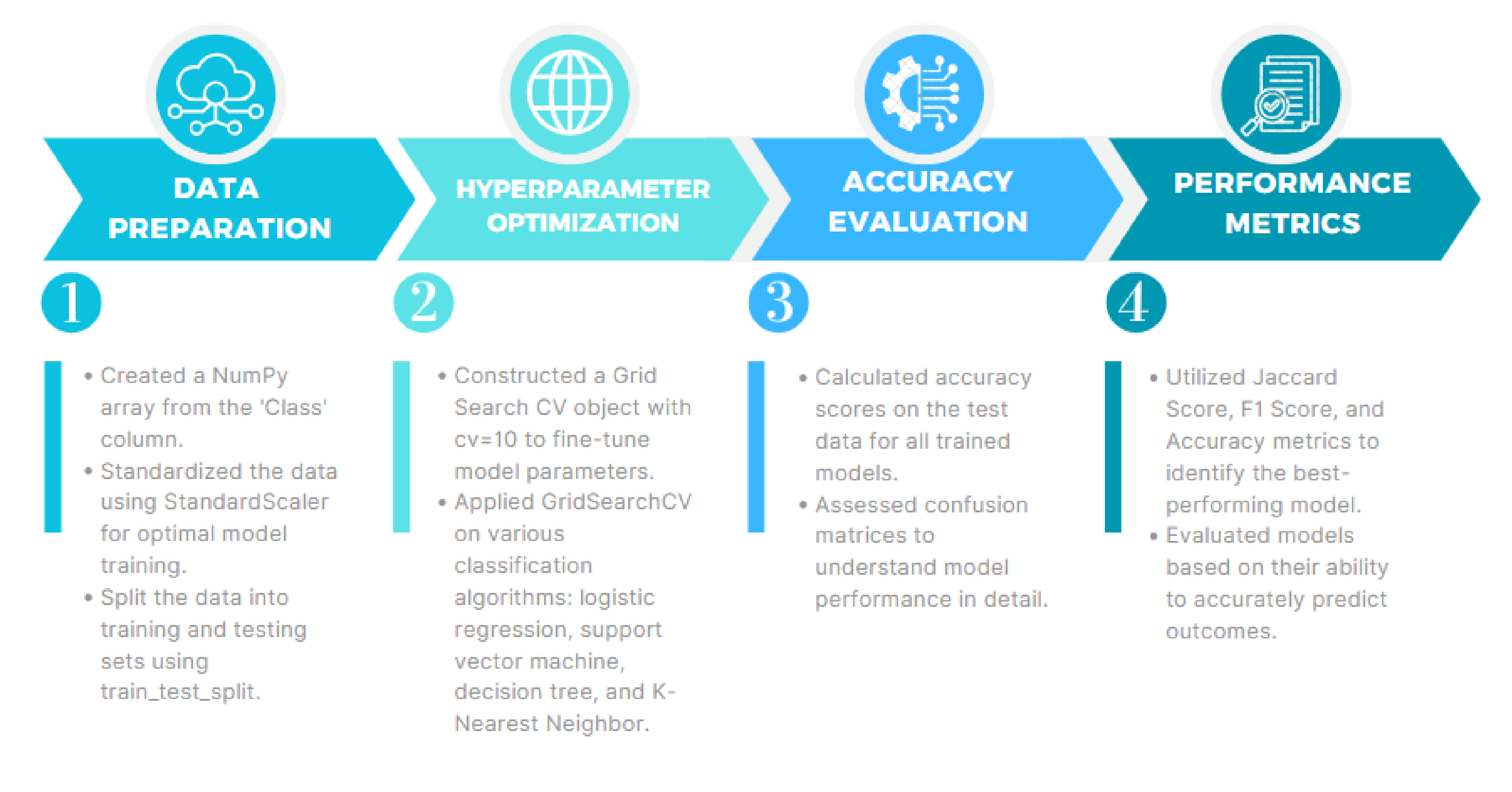
Build a Dashboard with Plotly Dash

- **Dropdown List - Launch Sites:**
  - Interaction: Select all or specific launch site.
  - Purpose: Tailored focus on site-specific or overall data.
- **Pie Chart - Successful Launches:**
  - Interaction: Display success/ unsuccessful percentages by site.
  - Purpose: Quick insight into success rates per site.
- **Slider - Payload Mass Range:**
  - Interaction: Set payload mass range.
  - Purpose: Explore success patterns within payload ranges.
- **Scatter Chart - Payload vs. Success:**
  - Interaction: Analyze payload's impact on success, by version.
  - Purpose: Uncover payload's influence on success rates.

Incorporating these elements provides an intuitive and interactive dashboard for comprehensive Falcon 9 launch analysis.

GitHub : Dashboard with Plotly Dash

Predictive Analysis (Classification)



GitHub URL: Space-X Machine Learning Prediction

Results

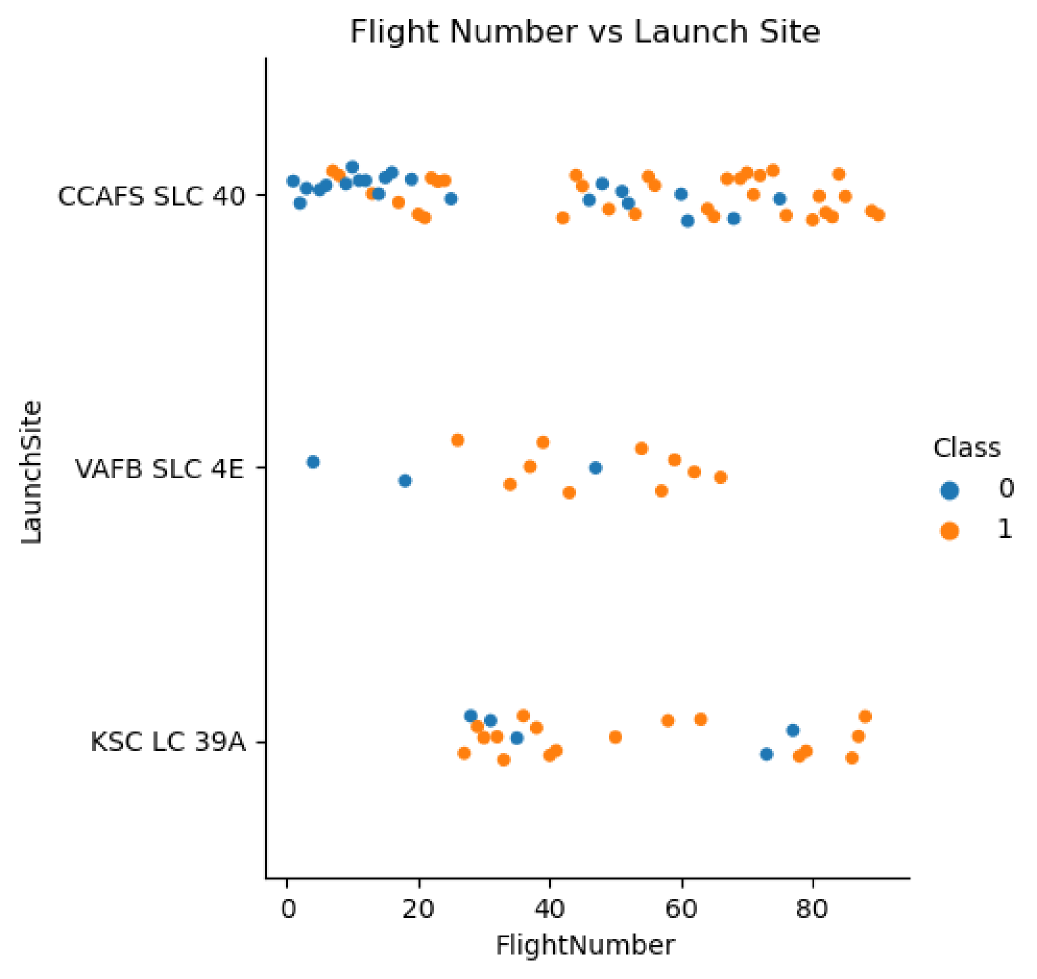
- Exploratory Data Analysis:
  - Launch success has shown improvement over time.
  - KSC LC-39A boasts the highest success rate among landing sites.
  - Orbits ES-L1, GEO, HEO, and SSO achieve 100% success.
- Visual Analytics:
  - Launch sites are strategically located near the equator and coastlines.
  - Sites maintain safe distances from potential damage zones while facilitating logistics.
- Predictive Analytics:
  - The Decision Tree model excels as the best predictive choice for the dataset.

These insights offer a holistic understanding of Falcon 9 launch dynamics, enabling informed strategies and decisions.

Section 2

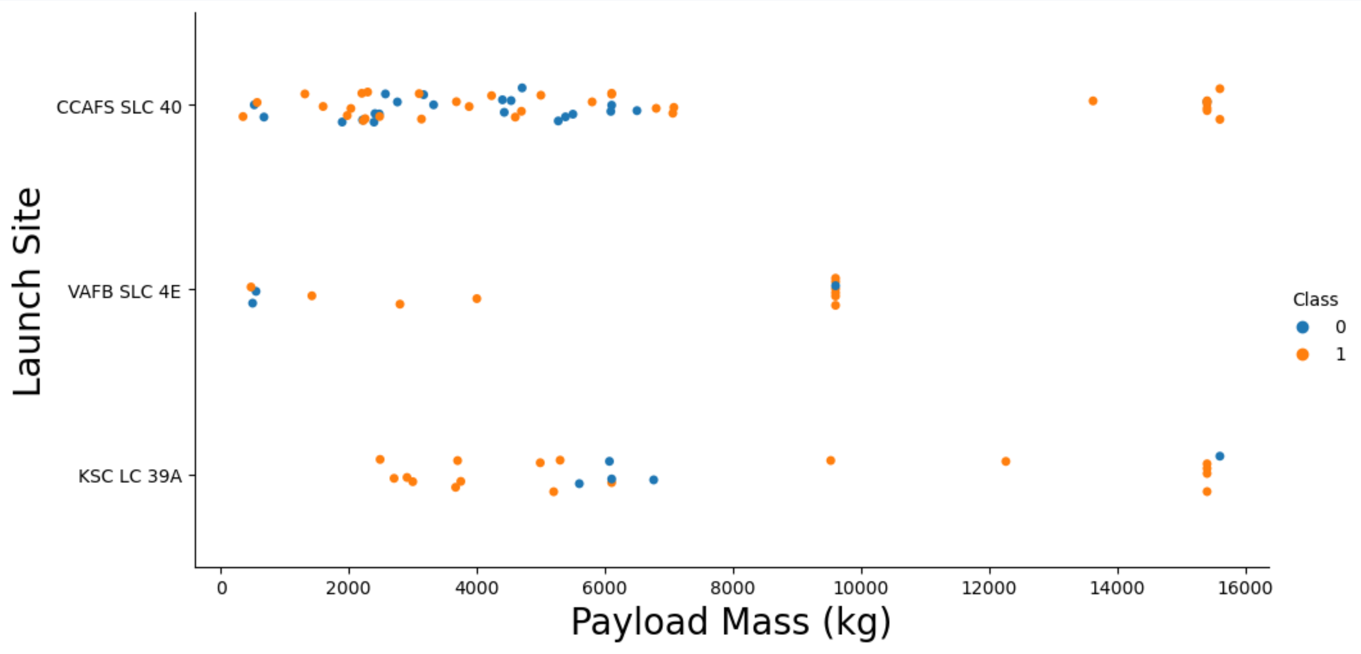
Insights drawn from EDA

Flight Number vs. Launch Site



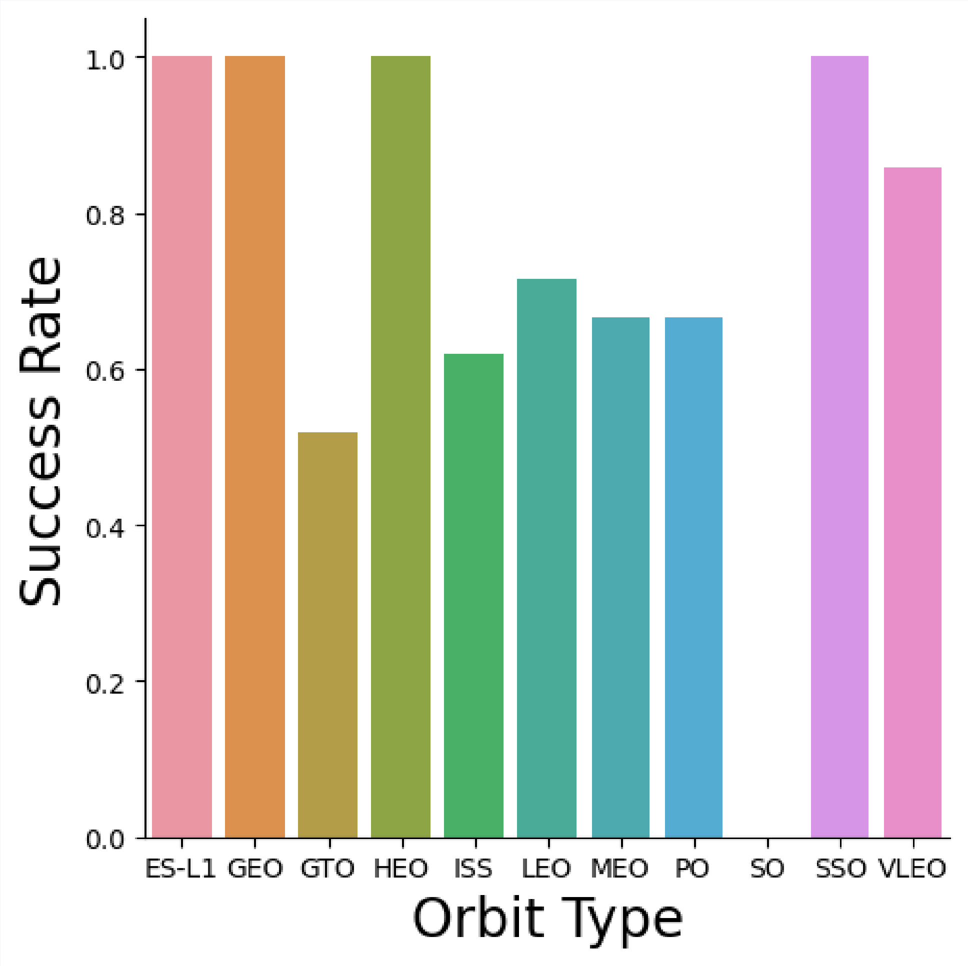
- This visual indicates a rise in success rate with the growing number of flights.
- Successful launches are depicted by blue dots, while unsuccessful ones are marked by a orange dot.
- An observable surge in successful flights becomes evident post the 40th launch.

Payload vs. Launch Site



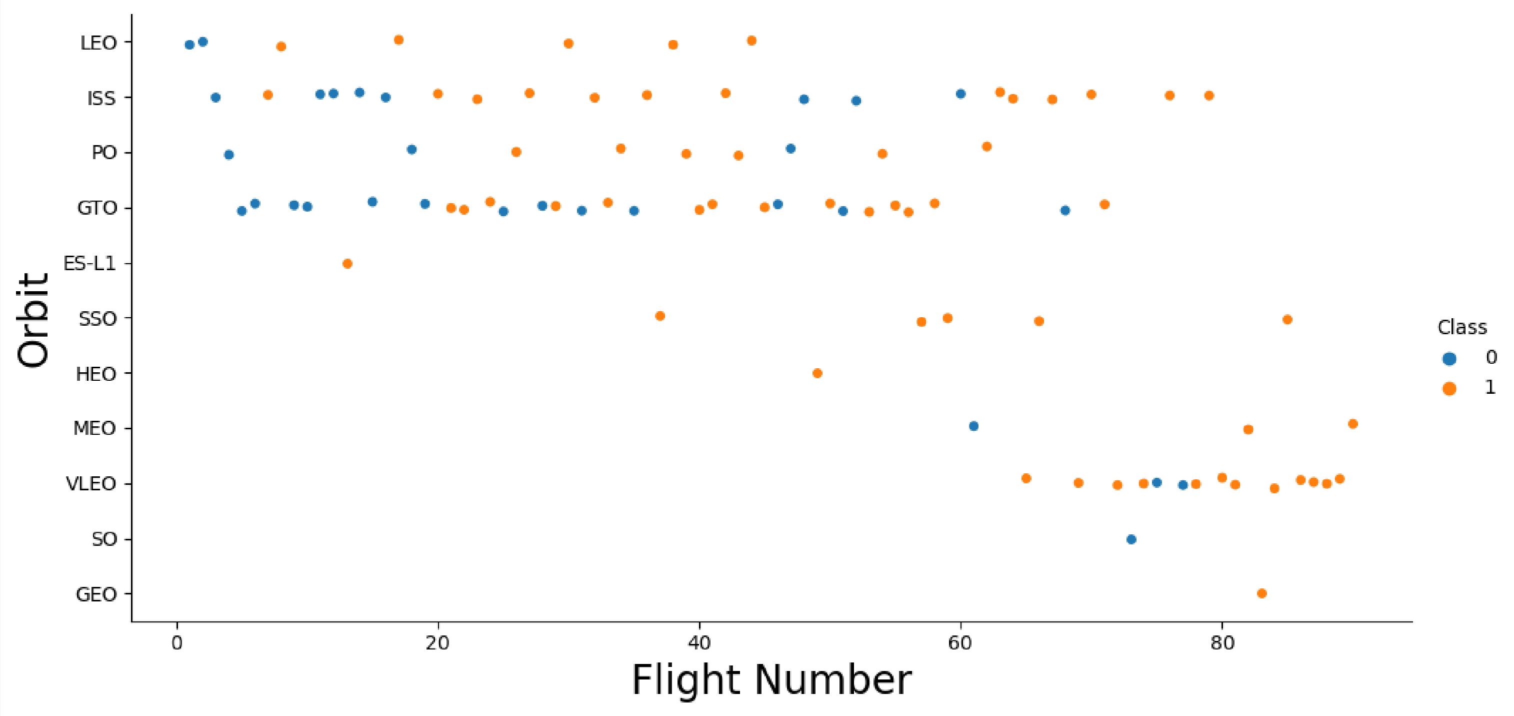
- In general, as the payload mass (kg) increases, the success rate tends to rise.
- Successful launches are prominent for payloads exceeding 7,000 kg.
- KSC LC 39A boasts a perfect 100% success rate for launches under 5,500 kg.
- VAFB SKC 4E has yet to launch payloads surpassing around 10,000 kg.

Success Rate vs. Orbit Type



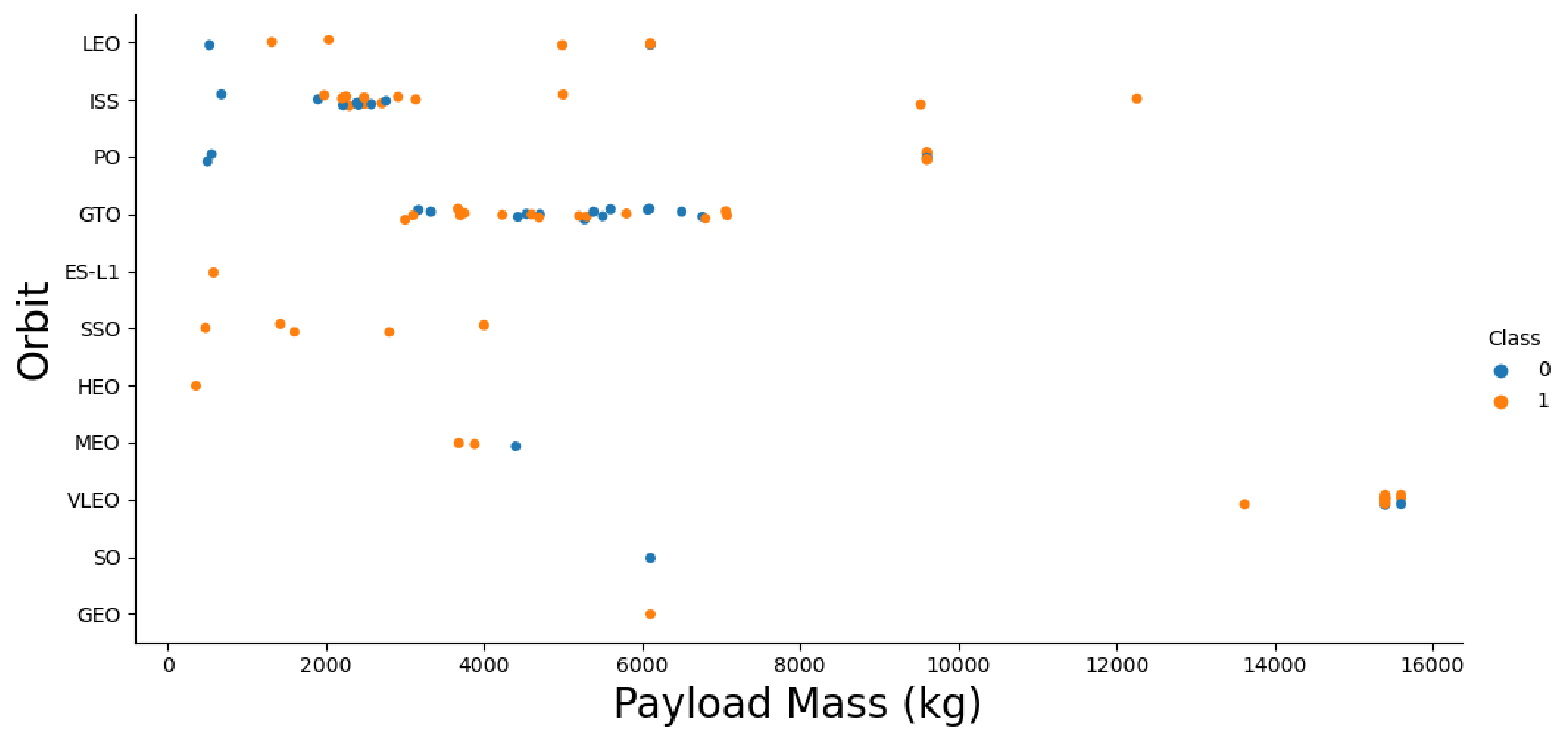
- Orbits with 100% Success: ES-L1, GEO, HEO, and SSO
- Orbits with 50%-80% Success: GTO, ISS, LEO, MEO, PO
- Orbits with 0% Success: SO

Flight Number vs. Orbit Type



- Generally, the success rate rises as the number of flights for each orbit increases.
- This pattern is particularly noticeable in the case of the LEO orbit.
- Interestingly, the GTO orbit does not conform to this trend.

Payload vs. Orbit Type

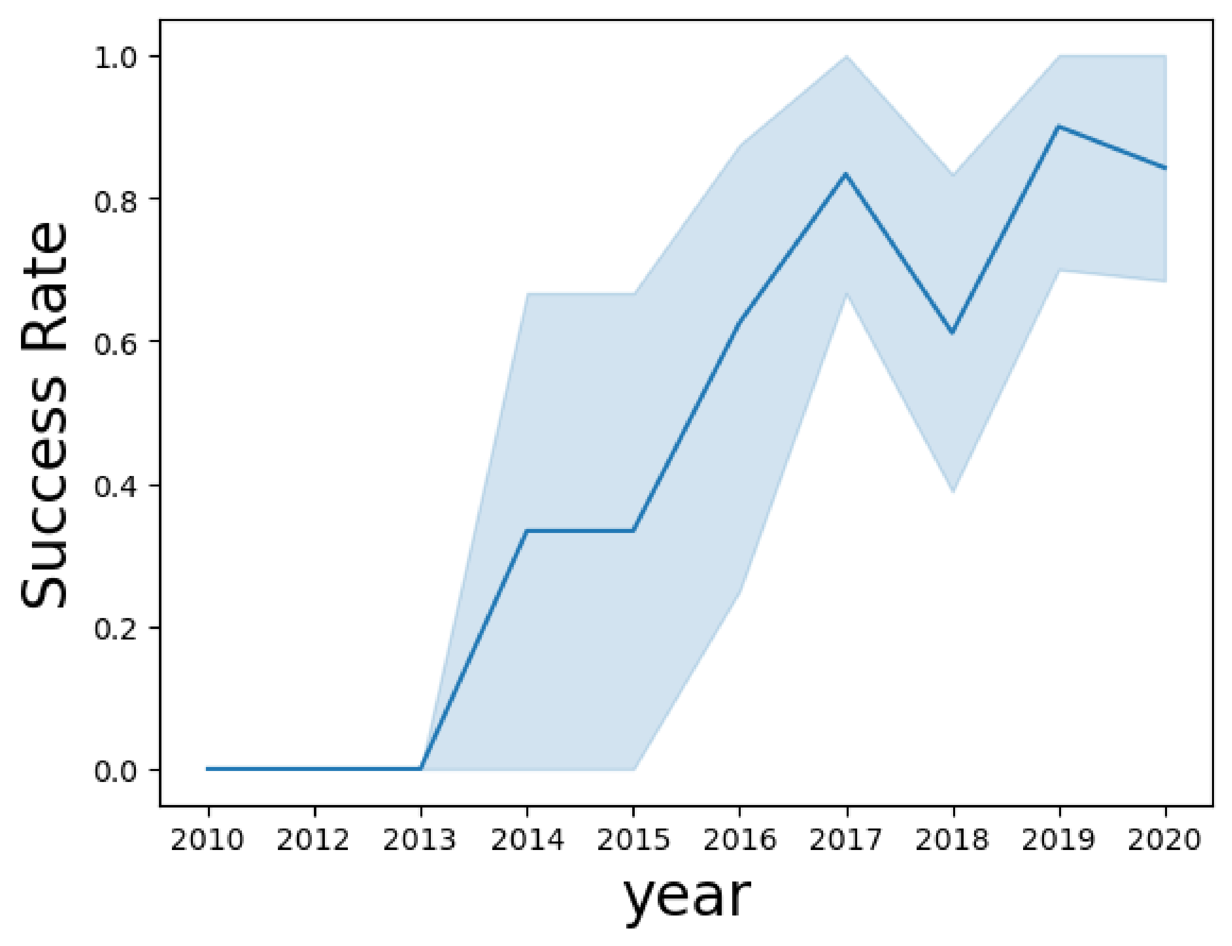


As payloads become heavier, success rates rise in the PO, SSO, LEO, and ISS orbits.

However, in the case of the GTO orbit, there isn't a clear correlation between orbit type and payload mass, as both successful and failed launches are evenly distributed.

Launch Success Yearly Trend

- Success rates saw enhancement from 2013-2017 and 2018-2019.
- Success rates declined from 2017-2018 and 2019-2020
- In summary, the success rate has demonstrated improvement since 2013.



All Launch Site Names

- Utilizing the DISTINCT clause, we retrieved solely the unique rows from the launch site column.
- The launch site names encompass:
  - CCAFS LC-40,
  - CCAFS SLC-40,
  - KSC LC-39A
  - VAFB SLC-4E.

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

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

To showcase solely the initial five outcomes where the launch_site name commences with 'CCA', the LIMIT and LIKE clauses were employed.

Total Payload Mass

Total Payload Mass

45,596 kg (total)

carried by boosters launched by NASA (CRS)

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Average Payload Mass

2,928 kg (average)

carried by booster version F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- By employing the MIN(DATE) function, we identified the date of the initial successful landing outcome on a ground pad
- Through the utilization of the WHERE clause, the outcomes were specifically filtered to instances where the 'landing_outcome' column indicated 'Success (ground pad)'.
- First Successful Ground Landing Date is "2015-12-22".

```
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters that effectively landed on a drone ship while possessing a payload mass exceeding 4000 but below 6000 is:
  - JCSAT-14
  - JCSAT-16
  - SES-10
  - SES-11 / EchoStar 105

```
%%sql
SELECT Payload
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(*) AS total_number
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

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

- 1 Failure in Flight
- 99 Success
- 1 Success (Payload status unclear)

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Employing the MAX() function within a subquery, we retrieved a compilation of boosters that have transported the highest payload mass.

2015 Launch Records

- Through the employment of the SELECT statement, various columns were extracted from the table. Additionally, the YEAR(DATE) function was harnessed to retrieve exclusively those rows marked with a launch date in the year 2015.

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The COUNT() function tallied distinct landing outcomes.
- Employing the WHERE and BETWEEN clauses, we narrowed down the outcomes to encompass only the time frame from June 4, 2010, to March 20, 2017.
- The GROUP BY clause organized the counts based on their respective outcomes.
- Finally, the ORDER BY and DESC clauses were implemented to arrange the outcomes in descending order.

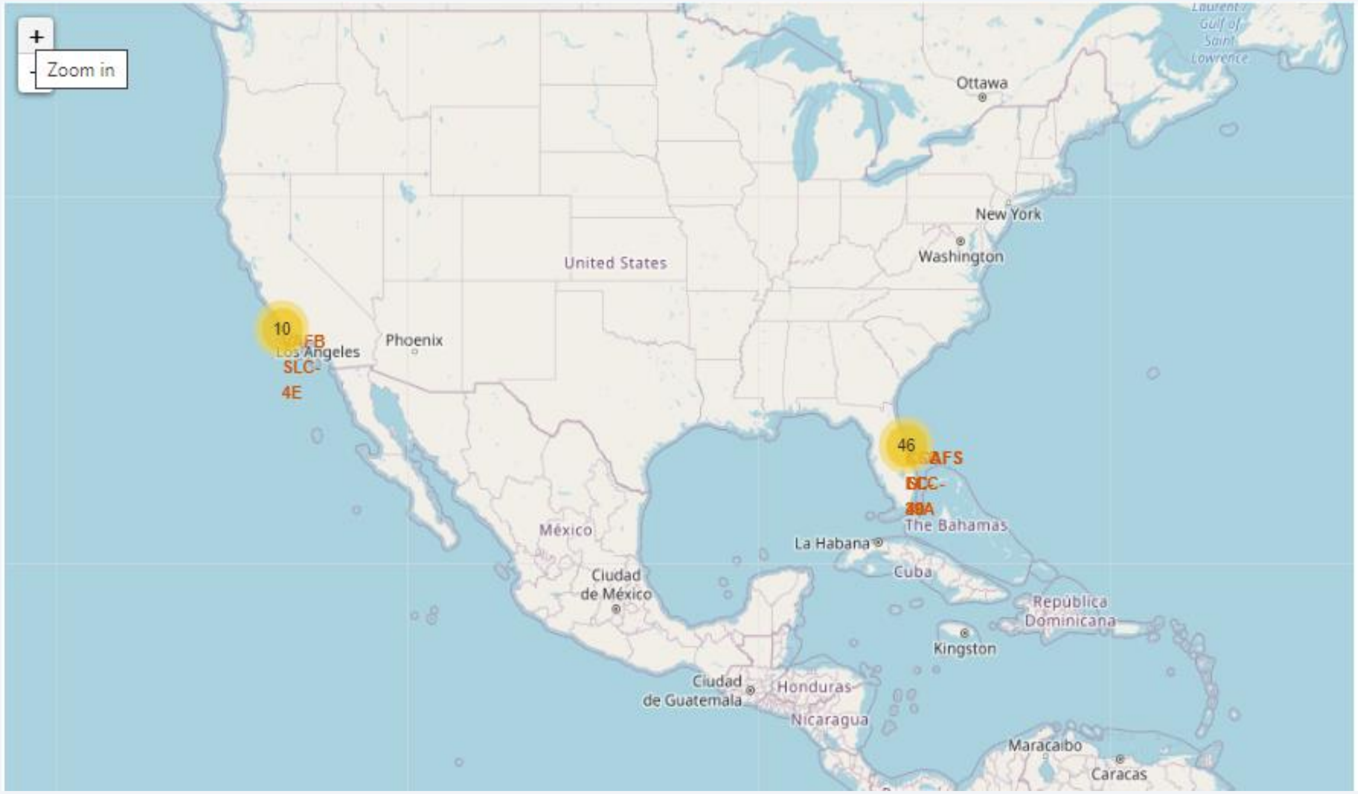
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

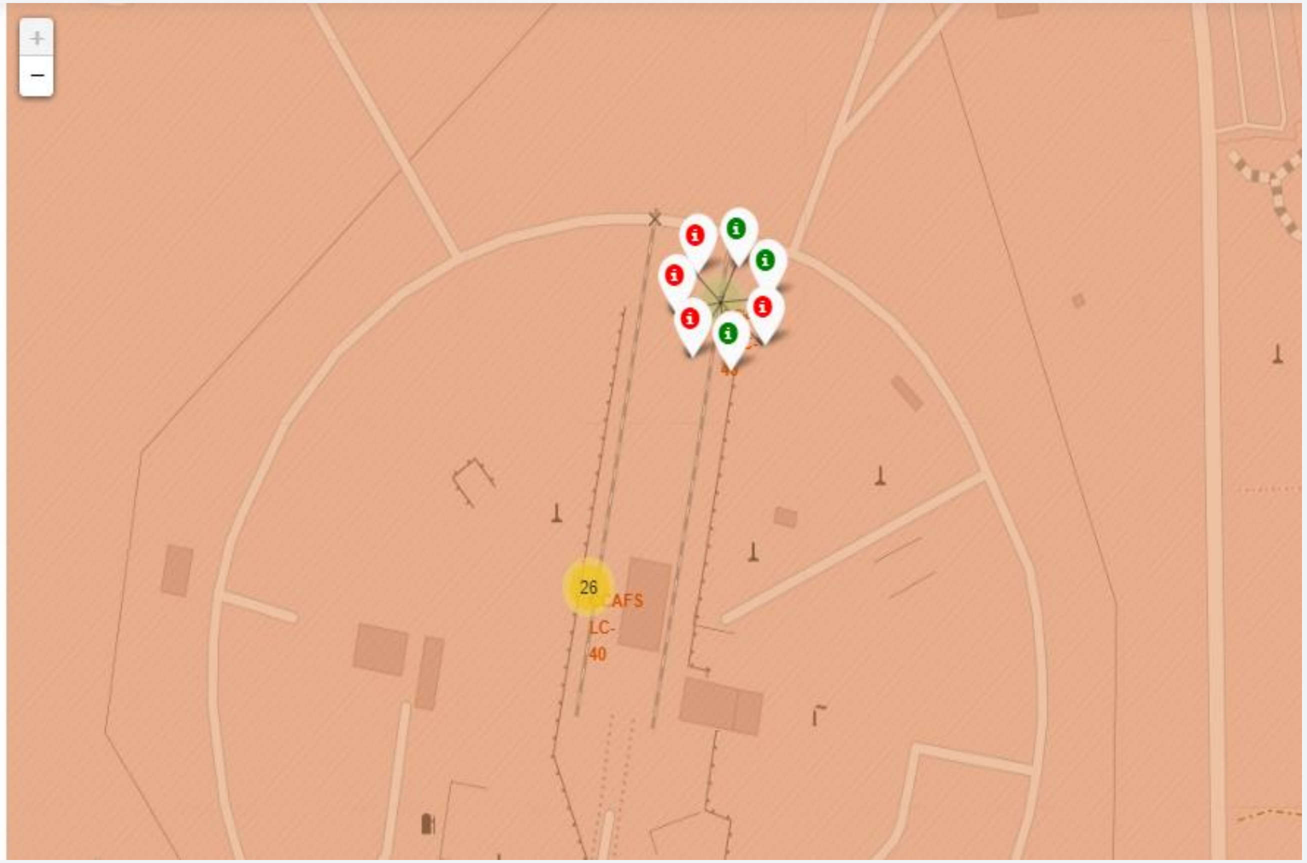
SpaceX Launch Sites Locations

- Yellow markers serve as indicators denoting the positions of all SpaceX launch sites within the US.
- These launch sites have been strategically positioned in proximity to coastal areas.



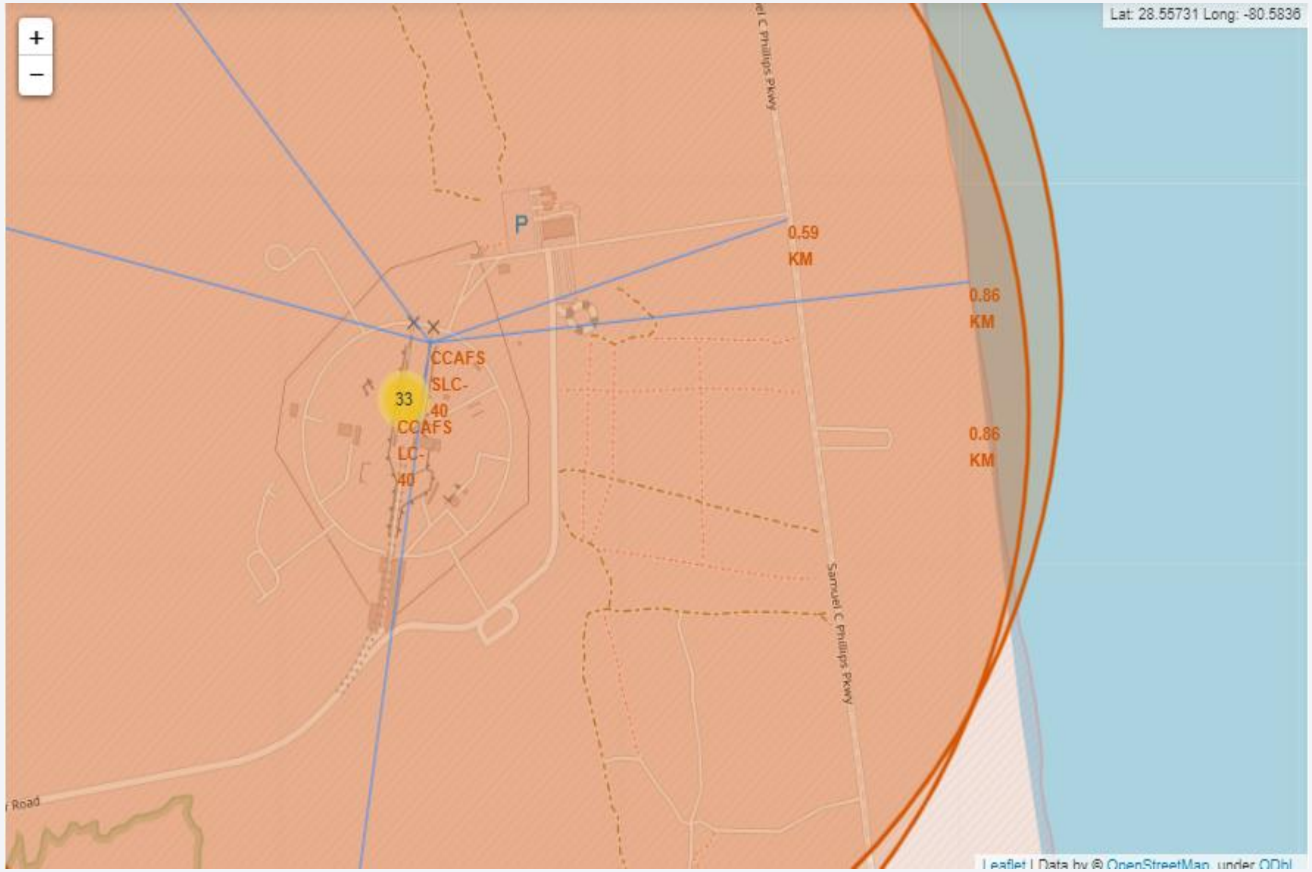
Success or Failure???

- Upon zooming in on a launch site, clicking on the site triggers the display of marker clusters.
- These clusters differentiate between successful landings (green) and unsuccessful ones (red).



<Folium Map Screenshot 3>

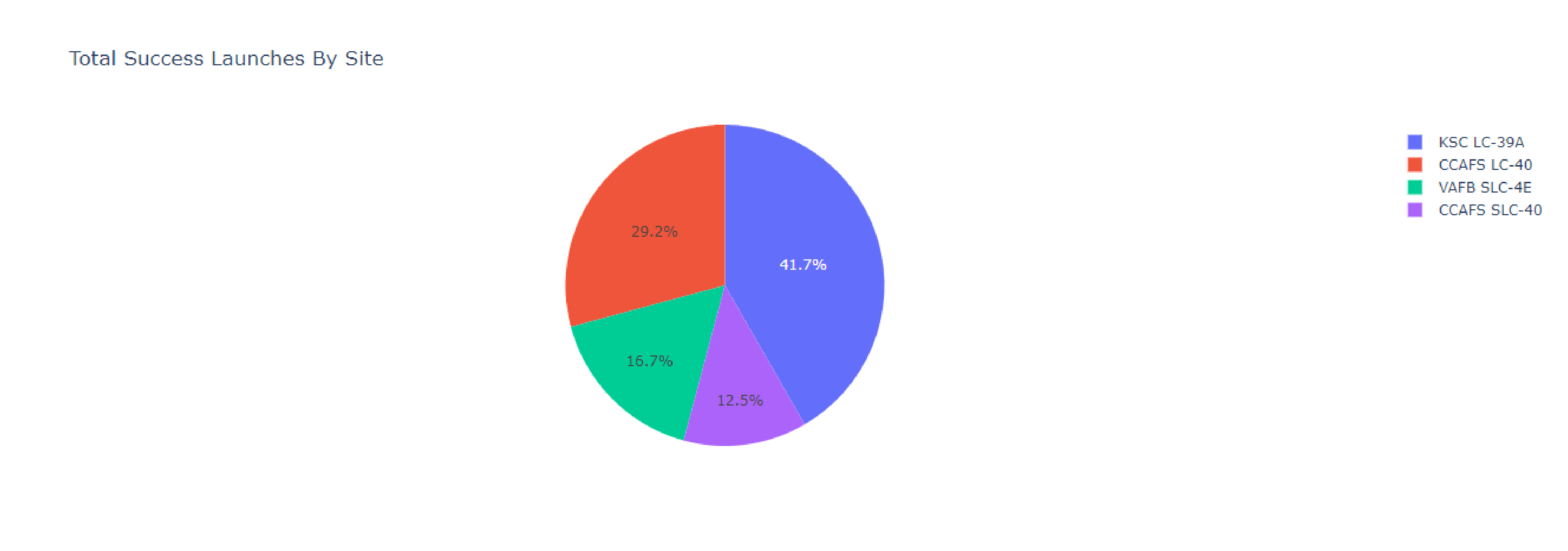
- The generated map indicates the proximity of the chosen launch site to a nearby highway, facilitating efficient transportation of personnel and equipment. Additionally, the launch site's closeness to coastlines is advantageous for conducting launch failure testing.
- Furthermore, the launch sites are intentionally situated at a distance from cities to ensure safety. (Accessible in the notebook for detailed viewing).



Section 4

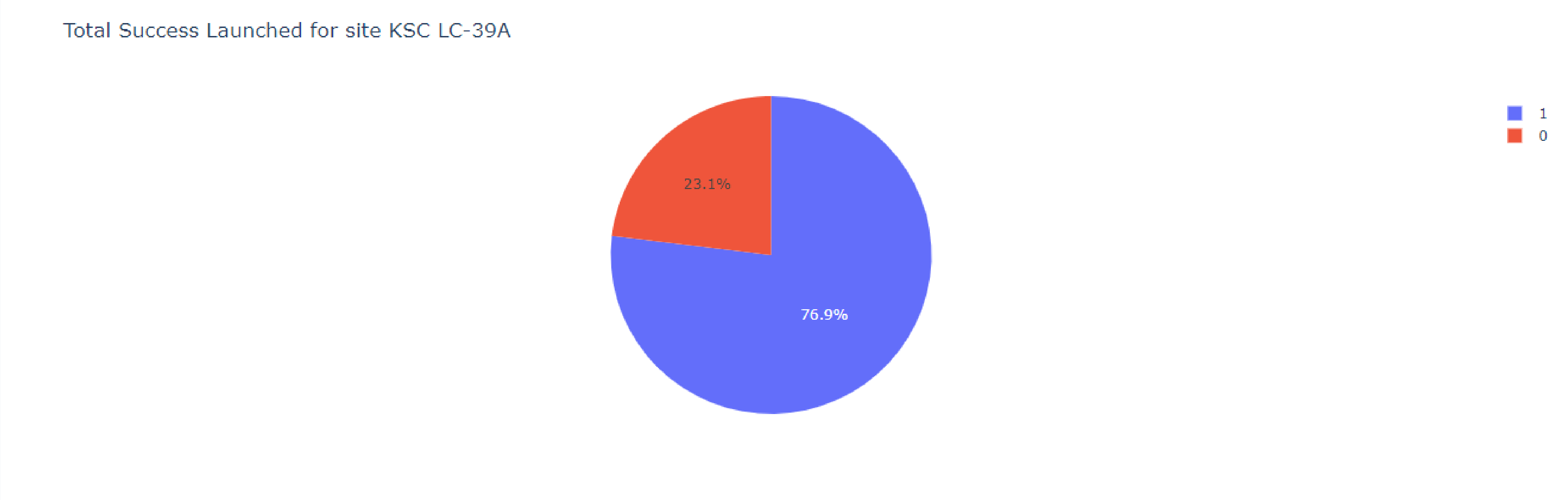
Build a Dashboard with Plotly Dash

Total Successful Launches By Site



- The KSC LC-39A Launch site has recorded the highest number of successful launches, totaling 10.

Launch Site Boasting the Highest Success Ratio

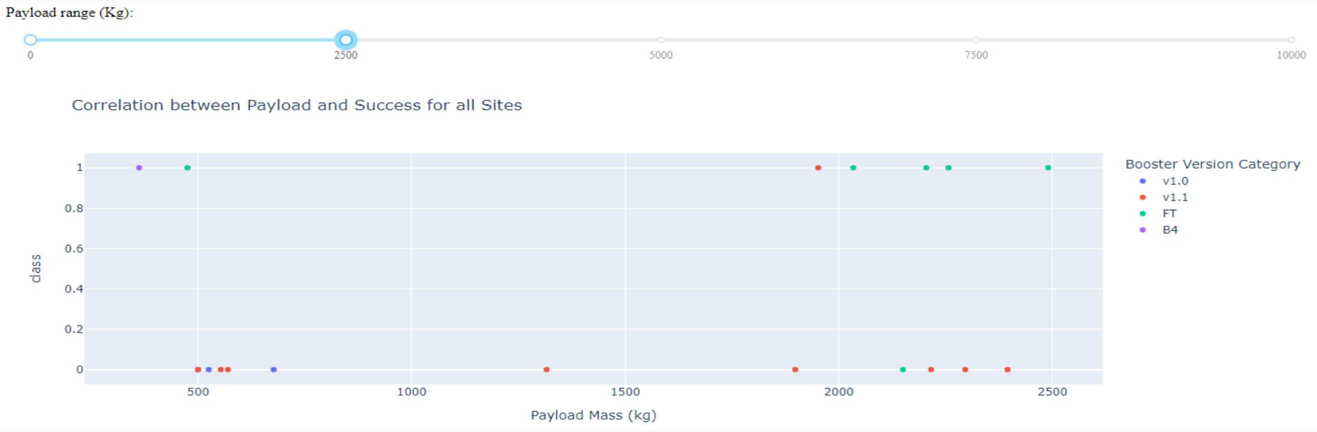


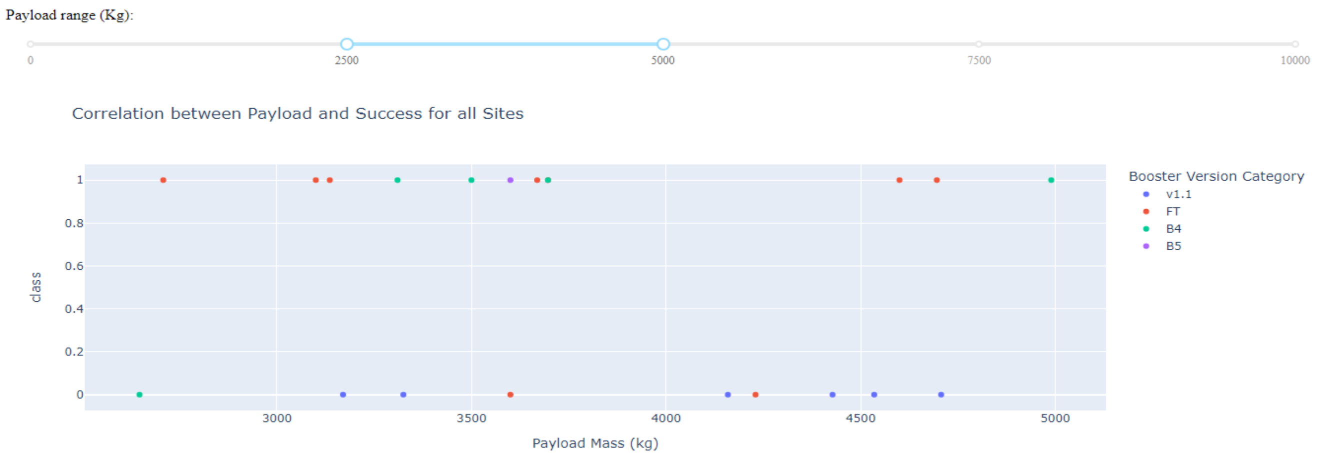
- The KSLC-39A exhibits the most noteworthy success rate, standing at 76.9%.

Payloads versus Launch Outcome

- The success rate for launches carrying payloads of 0-2500 kg is slightly below that of payloads ranging from 2500 to 5000 kg. Interestingly, the disparity between the two is minimal.

- Notably, the booster version v1.1 demonstrates the highest success rate across both weight ranges.



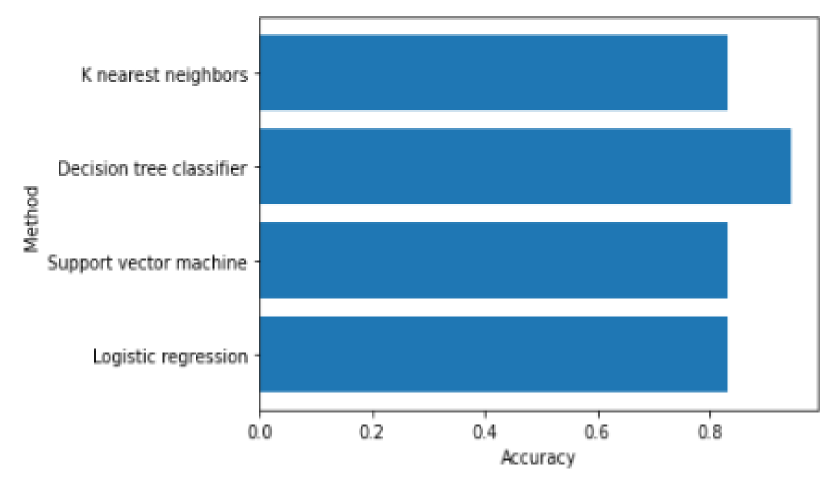


Section 5

Predictive Analysis (Classification)

Classification Accuracy

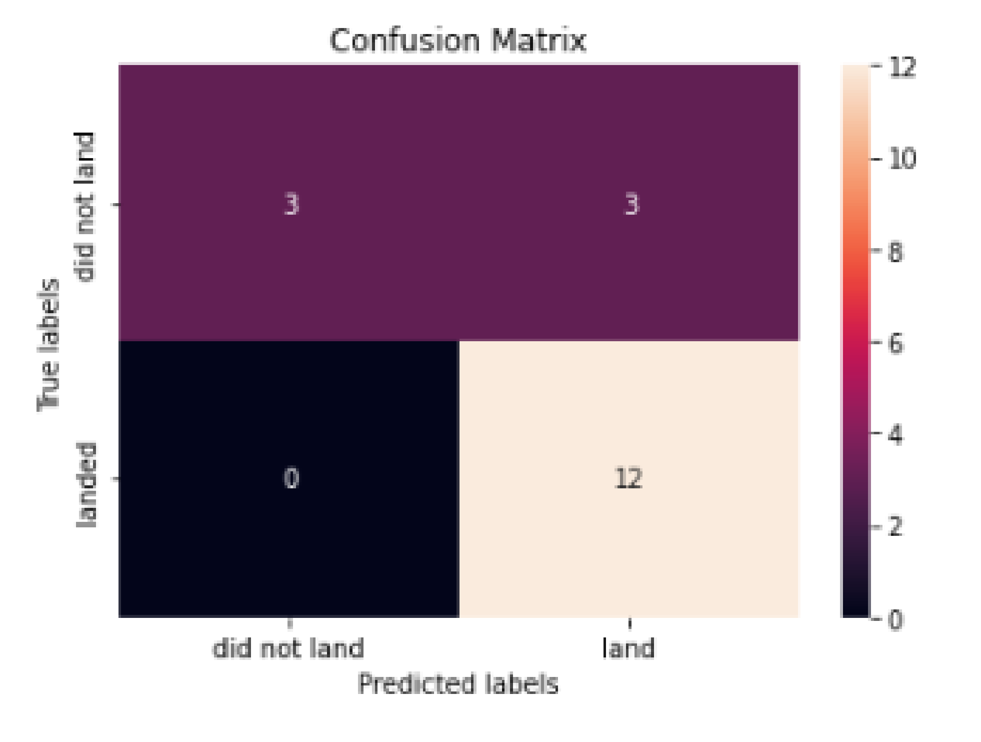
- The Decision Tree classifier achieved the highest accuracy, reaching an impressive 94%.



	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333

Confusion Matrix

- The model accurately predicted 12 successful landings when the actual label was successful (True Positive), and correctly identified 3 unsuccessful landings when the actual label was a failure (True Negative).
- However, the model erroneously predicted 3 successful landings when the actual label was an unsuccessful landing (False Positive).
- Overall, the model demonstrated a tendency to predict successful landings.



Conclusions

- A positive correlation between the number of flights and success rate, with an observable improvement over the years.
- Notable success rates in specific orbits, including SSO, HEO, GEO, and ES-L1.
- A relationship between success rate and payload mass, favoring lighter payloads.
- Strategic launch site placement near transportation routes and away from urban areas.
- The Decision Tree Classifier as the optimal predictive model, boasting an impressive 94% accuracy.

Appendix

- Coursera Project Link: https://www.coursera.org/learn/applied-data-science-capstone/home/welcome
- GitHub Repository: https://github.com/withouttheh/IBM-Data-Science-Capstone-Project

Thank you!