IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Manuel Parra
January 13 2025

Outline

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

Executive Summary

Methodologies

1. Data Collection and Preparation
2. Exploratory Data Analysis (EDA)
3. Model Training and Evaluation

Importance

- Efficient Planning: *Accurate predictions enable better planning of resources and staff, ensuring that reusable boosters are effectively utilized.*
- Customer Satisfaction: *Correctly identifying successful launches and potential failures improves expectation management and reduces customer dissatisfaction.*
- Resource Optimization: *Minimizing false negatives can lead to better resource utilization and more efficient logistics.*
- Informed Decision-Making: *Evaluation metrics provide valuable insights for making informed decisions about model improvement and process optimization.*

Introduction

Project Background and Context

SpaceX has revolutionized the aerospace industry by significantly reducing the cost of rocket launches through the reusability of its Falcon 9 first-stage boosters. The ability to reuse these boosters allows SpaceX to offer launches at a cost of 62 million dollars, compared to other providers that charge upwards of 165 million dollars. This cost efficiency is a key factor in SpaceX's competitive advantage.

Problems to Address

The primary objective of this project is to predict whether the Falcon 9 first stage will successfully land after launch. Accurately predicting landing success can help determine the cost of a launch and provide valuable insights for alternate companies looking to compete with SpaceX. By analyzing various factors such as payload mass, orbit type, and launch site location, we aim to build a robust machine learning model that can reliably predict landing outcomes.

This project involves performing exploratory data analysis (EDA), feature engineering, and model training to identify the best-performing classification model. The findings from this analysis can be used to optimize launch strategies and improve the overall success rate of rocket launches.

Section 1

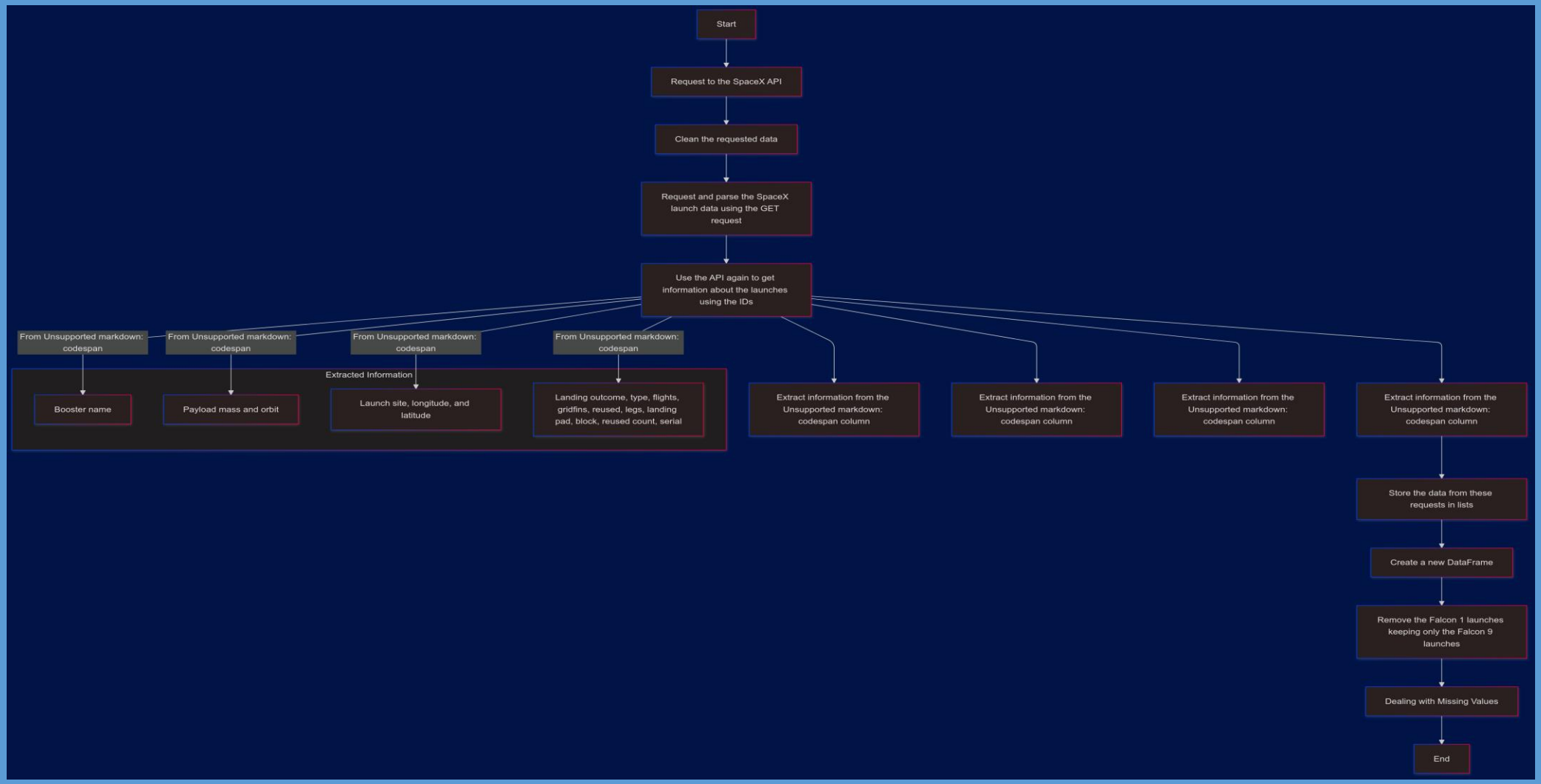
Methodology

Methodology

Executive Summary

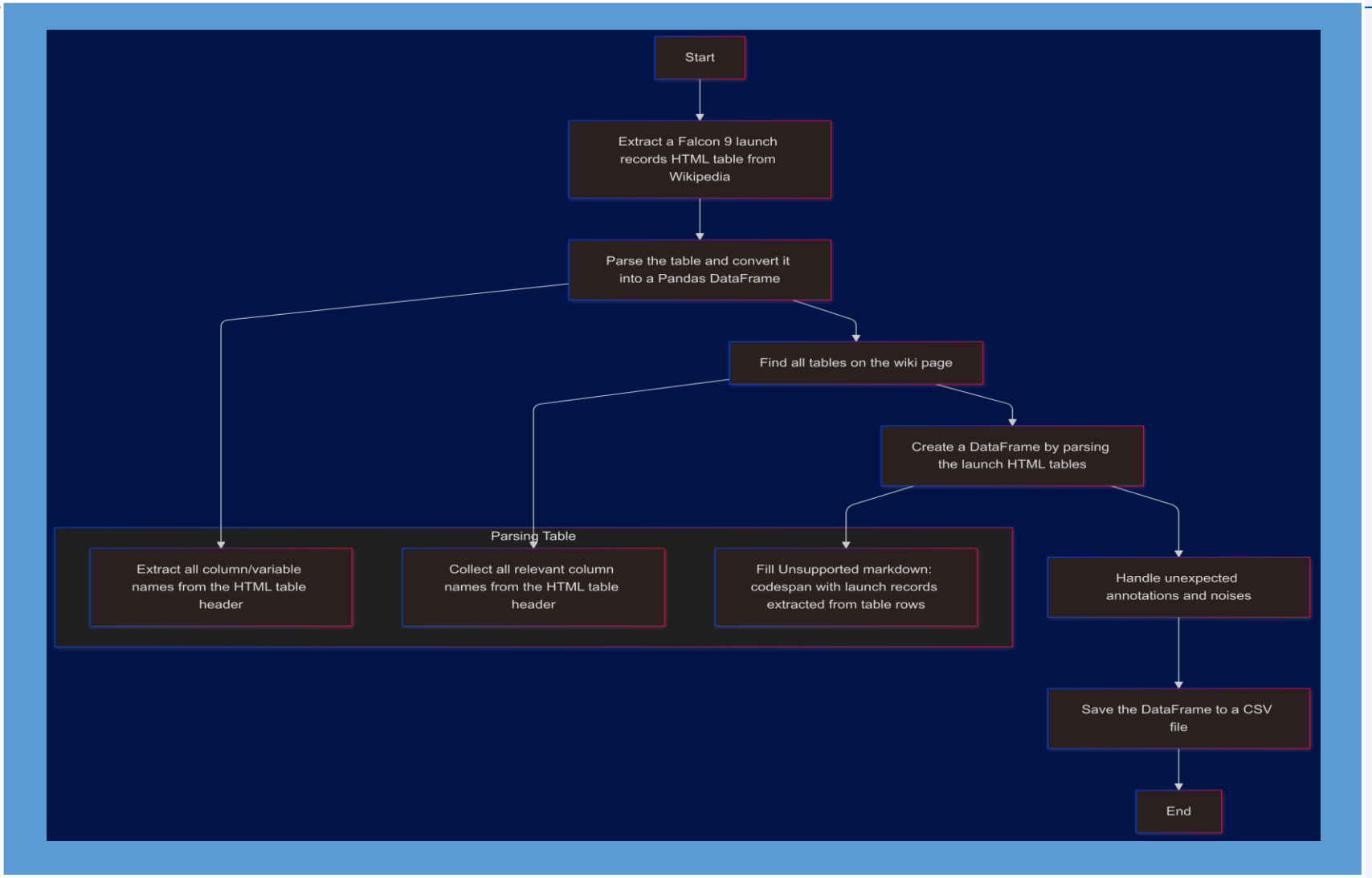
- **Data collection methodology:**
  - Describe how data was collected
- **Perform data wrangling**
  - Describe how data was processed
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API



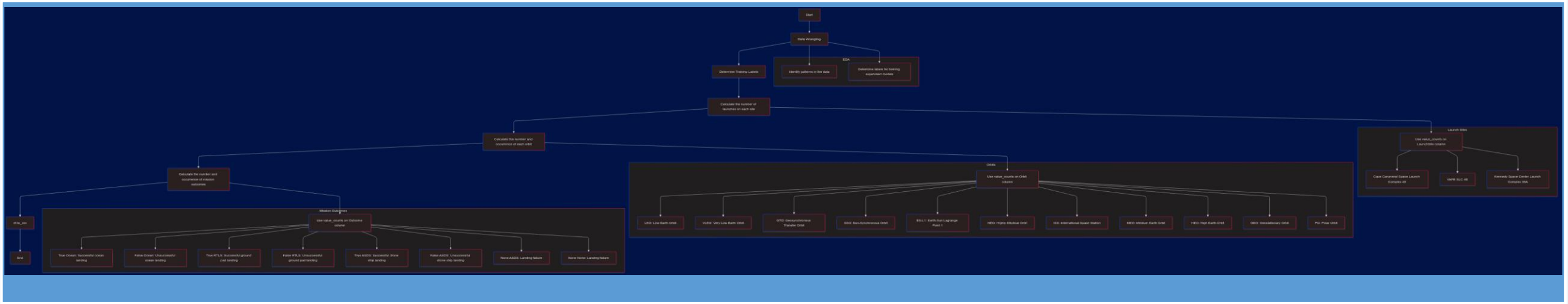
- Consult the **Notebook here**

Data Collection - Scraping



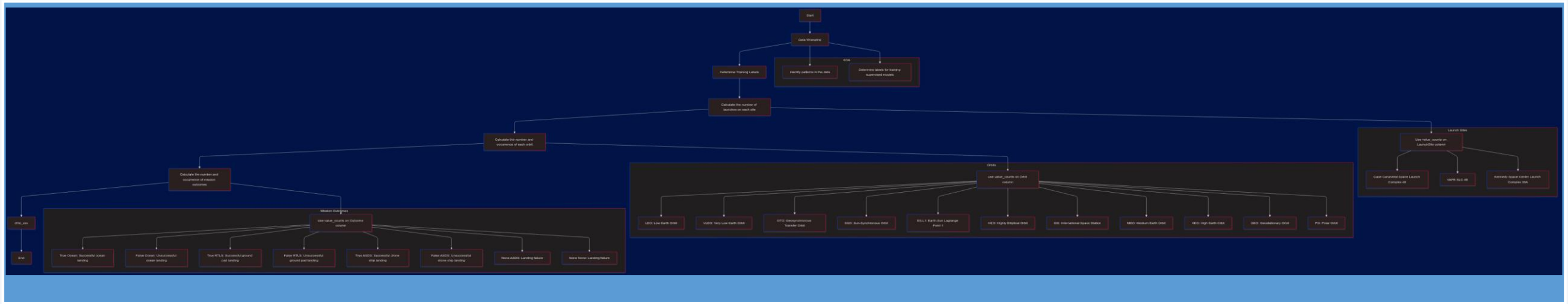
- Consult the **Notebook here**

Data Wrangling



1. **Start**
   - Begin the data wrangling process.
2. **Perform Exploratory Data Analysis (EDA):**
   - Identify patterns in the data.
   - Determine labels for training supervised models.
3. **Determine Training Labels:**
   - Convert mission outcomes into binary labels for model training:
     - `1` means the rocket successfully landed.
     - `0` means the landing was unsuccessful.

Data Wrangling



4. **Calculate the Number of Launches on Each Site:**
   - Use the `value_counts()` method on the `LaunchSite` column to determine the number of launches at each site:
     - Cape Canaveral Space Launch Complex 40
     - VAFB SLC 4E
     - Kennedy Space Center Launch Complex 39A
5. **Calculate the Number and Occurrence of Each Orbit**
6. **Calculate the Number and Occurrence of Mission Outcomes**
   - Use the `value_counts()` method on the `Outcome` column to determine the number of mission outcomes:
7. **Save the Processed Data to a CSV File**
   - Use `df.to_csv("dataset_part_2.csv", index=False)` to save the cleaned and processed data.
8. **End**
   - Conclude the data wrangling process.

- Consult the **Notebook here**

EDA with Data Visualization

1. Flight Number vs. Payload Mass (payloadVSnflights.png):
   - **Purpose:** To observe how the `FlightNumber` (indicating continuous launch attempts) and `PayloadMass` variables affect the launch outcome.
   - **Chart Used:** Scatter plot of `FlightNumber` vs. `PayloadMass` with the launch outcome overlaid.
2. Flight Number vs. Launch Site (task2.png):
   - **Purpose:** To visualize the relationship between `FlightNumber` and `LaunchSite`.
   - **Chart Used:** `catplot` with `FlightNumber` on the x-axis, `LaunchSite` on the y-axis, and `class` (outcome) as the hue.
3. Payload vs. Launch Site (task3.png):
   - **Purpose:** To observe if there is any relationship between launch sites and their payload mass.
   - **Chart Used**: Scatter plot of `Payload` vs. `LaunchSite`.
4. Success Rate of Each Orbit Type (task3.png):
   - **Purpose**: To visually check if there is any relationship between success rate and orbit type.
   - **Chart Used:** Bar chart for the success rate of each orbit.
5. Flight Number vs. Orbit Type (task4.png):
   - **Purpose**: To see if there is any relationship between `FlightNumber` and `Orbit` type.
   - **Chart Used:** Scatter plot of `FlightNumber` vs. `Orbit`.
6. Payload vs. Orbit Type (task5.png):
   - **Purpose:** To reveal the relationship between `Payload` and `Orbit` type.
   - **Chart Used:** Scatter plot of `Payload` vs. `Orbit`.
7. Launch Success Yearly Trend (task6.png):
   - **Purpose:** To observe the yearly trend of launch success rates.
   - **Chart Used:** Line chart with `Year` on the x-axis and average success rate on the y-axis.

- Consult the **Notebook here**

EDA with SQL

- **Q1:** *What are the unique launch sites used by SpaceX?*
- **Q2:** *What are the first 5 records where the launch site name starts with 'CCA'?*
- **Q3:** *What is the total payload mass carried by boosters launched for NASA (CRS)?*
- **Q4:** *What is the average payload mass carried by the booster version F9 v1.1?*
- **Q5:** *When was the first successful landing on a ground pad achieved?*
- **Q6:** *Which boosters successfully landed on a drone ship with a payload mass between 4,000 kg and 6,000 kg?*
- **Q7:** *What is the total number of successful and failed missions?*
- **Q8:** *Which booster versions carried the maximum payload mass?*
- **Q9:** *What are the failure landing outcomes in drone ships during 2015, along with booster versions and launch sites?*
- **Q10:** *What are the counts of different landing outcomes between 2010-06-04 and 2017-03-20, ranked in descending order?*

- Consult the **Notebook here**

Build an Interactive Map with Folium

- **Mark all launch sites on a map**
  - **Object:** Markers and Circles
  - **Explanation:** Markers and circles were added to visually represent the locations of the launch sites on the map. This helps in identifying the geographical distribution of the launch sites.

- **Add circles for each launch site**
  - **Object:** Circles with Popups
  - **Explanation:** Circles with popups were added to highlight each launch site and provide additional information when hovered over. This enhances the visual representation and provides more context about each site.

- **Mark the success/failed launches for each site on the map**
  - **Object:** Markers with Clusters
  - **Explanation:** Markers were added to indicate the success or failure of each launch. This helps in visually analyzing which sites have higher success rates and identifying patterns related to launch outcomes.

- **Calculate the distances between a launch site to its proximities**
  - **Object:** MousePosition, Markers, and PolyLines
  - **Explanation:** The `MousePosition` object was added to interactively get coordinates on the map. Markers were used to indicate the distance between a launch site and its proximities, and `PolyLine` was used to visually connect the launch site to the selected proximity point. This helps in analyzing the proximity factors that may affect the launch success rate.

- Consult the **Notebook here**

Build a Dashboard with Plotly Dash

1. Launch Site Dropdown:

- **Purpose:** *To provide a user-friendly way to filter data by launch site.*
- **Benefit:** *Enhances the dashboard's interactivity and allows for more granular analysis of launch data by site.*

2. Pie Chart for Launch Success:

- **Purpose:** *To visually represent the success rates of launches.*
- **Benefit:** *Offers a quick and intuitive way to compare the success rates of different launch sites, aiding in the identification of the most successful sites.*

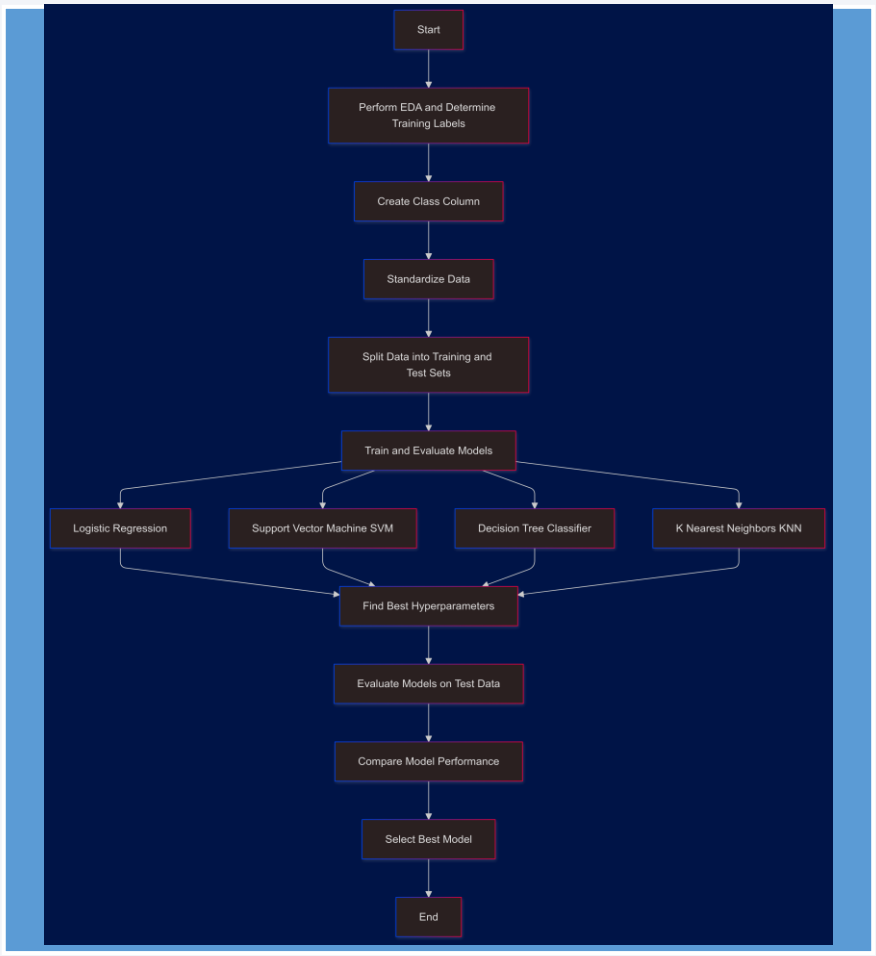
3. Range Slider for Payload:

- **Purpose:** *To enable users to select and analyze specific payload ranges.*
- **Benefit:** *Facilitates the exploration of how payload weight affects launch success, providing valuable insights into optimal payload ranges.*

4. Scatter Plot for Payload vs. Launch Outcome:

- **Purpose:** *To visualize the relationship between payload mass and launch outcomes.*
- **Benefit:** *Helps users identify patterns and trends in the data, such as which payload ranges and Booster versions are associated with higher success rates.*

Predictive Analysis (Classification)



- Create a NumPy array from the column `Class`

- **Action:** *Apply the method `to_numpy()` to the `Class` column in `data` and assign it to the variable `Y`.*
- **Result:** *`Y` is a Pandas series containing the class labels.*

- Standardize the data in `X`

- **Action:** *Use the `StandardScaler` to standardize the data in `X` and reassign it to the variable `X`.*
- **Result:** *`X` is standardized and ready for model training.*

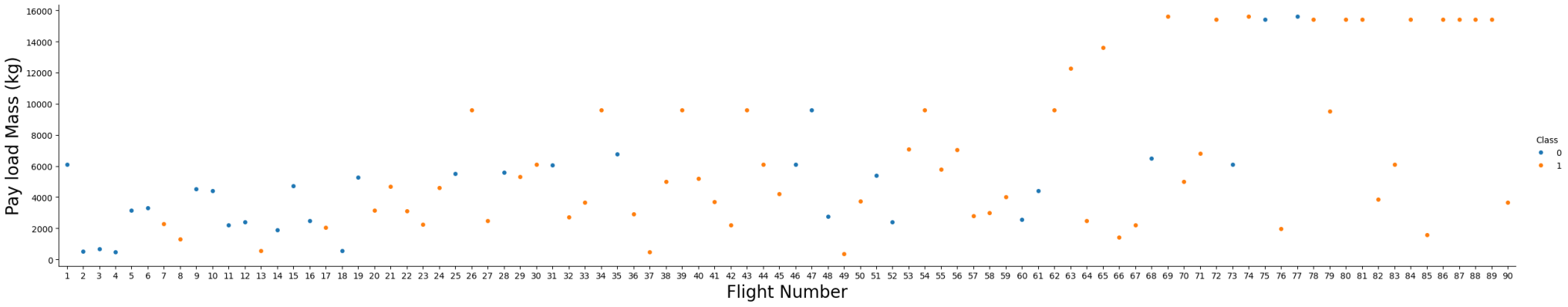
- Split the data into training and test sets

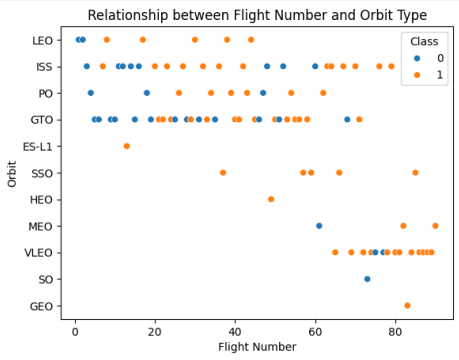
- **Action:** *Use the `train_test_split` function to split `X` and `Y` into training and test data with `test_size = 0.2` and `random_state = 2`.*

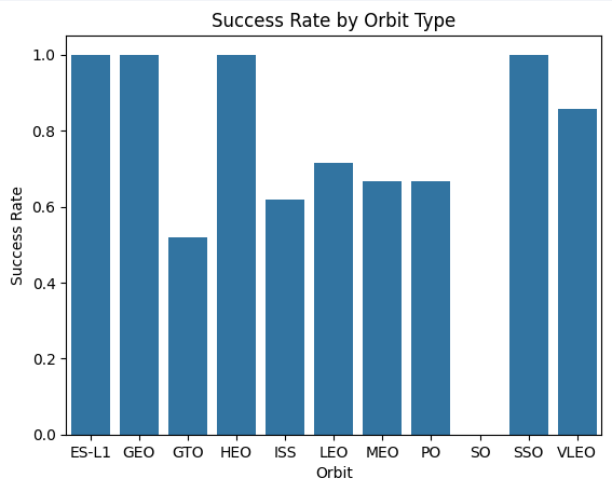
- Consult the **Notebook here**

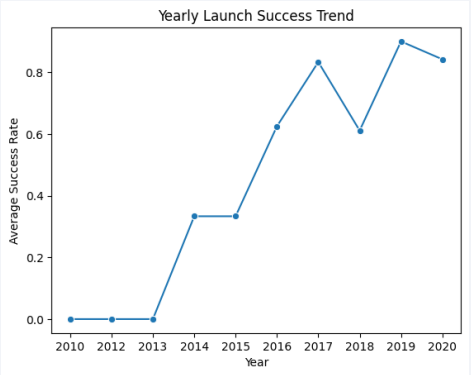
Results

- Exploratory data analysis results



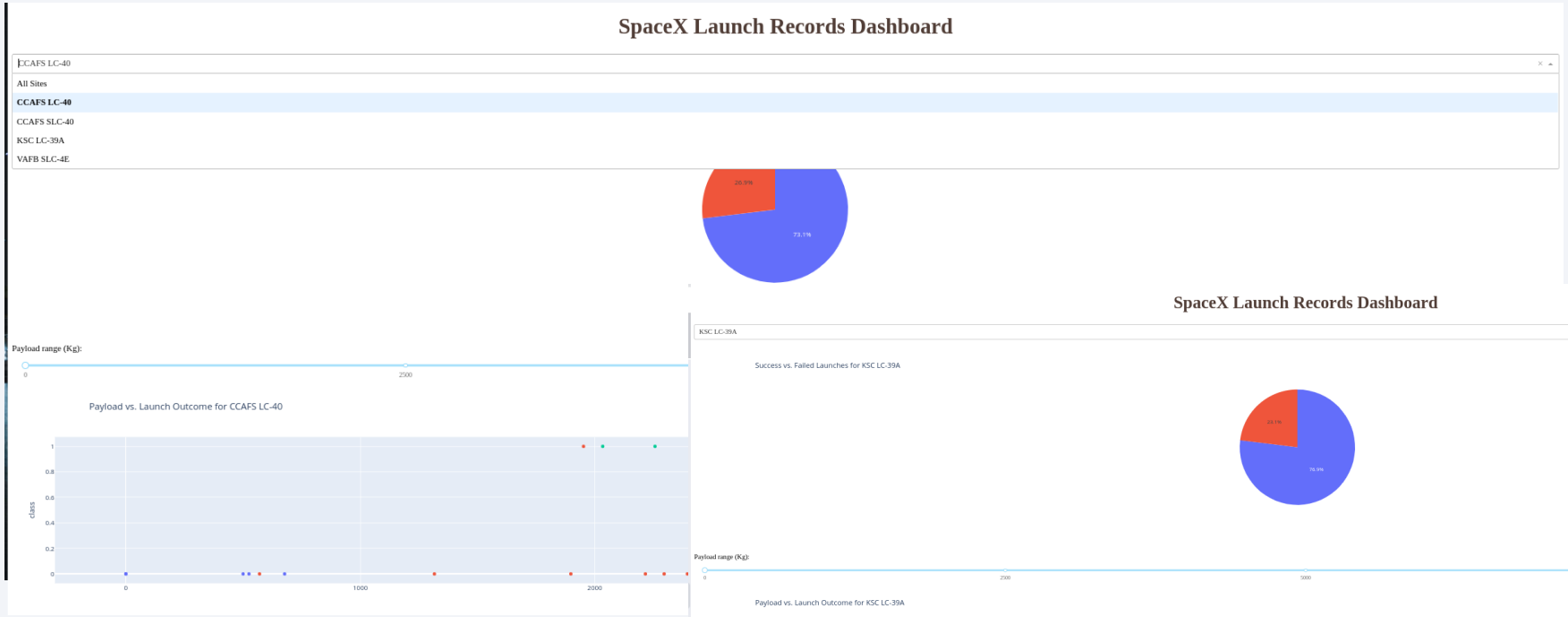


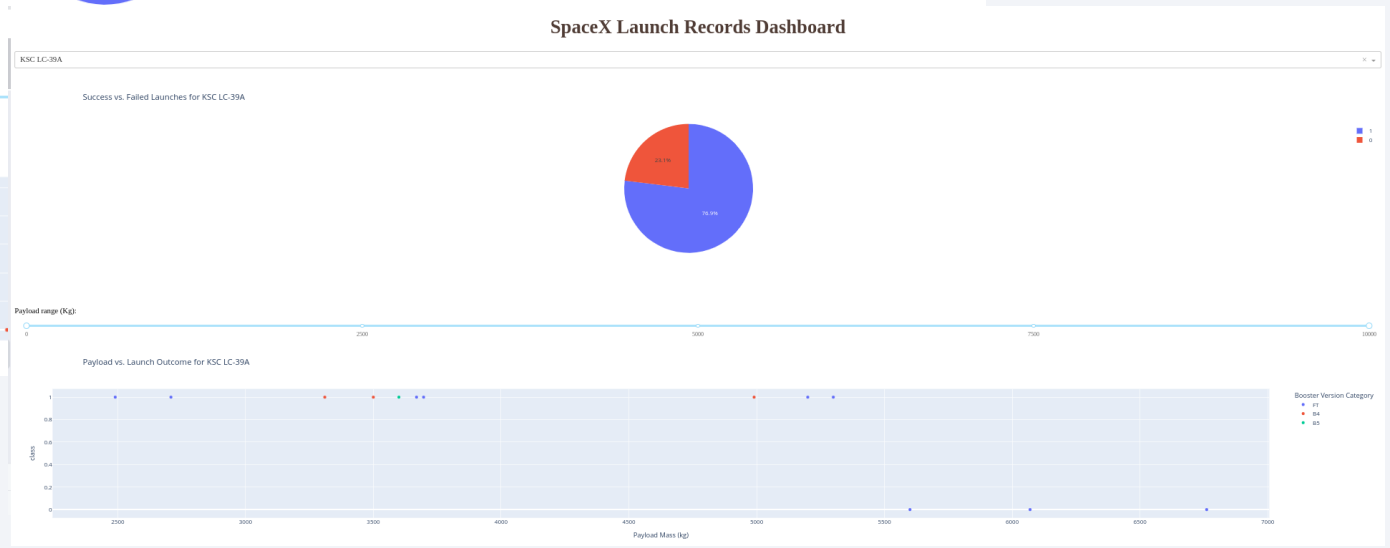




Results

- Interactive analytics demo in screenshots

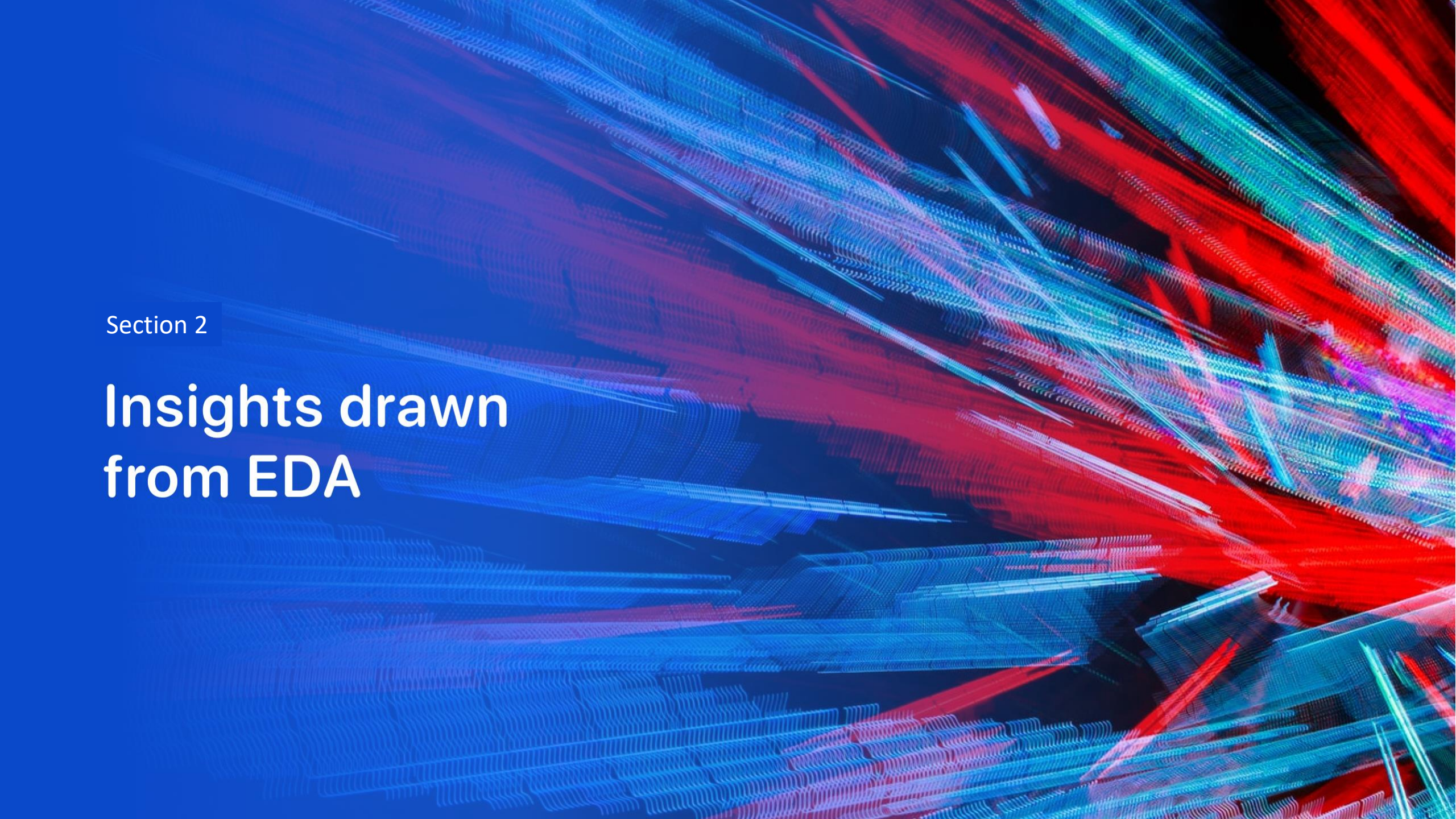




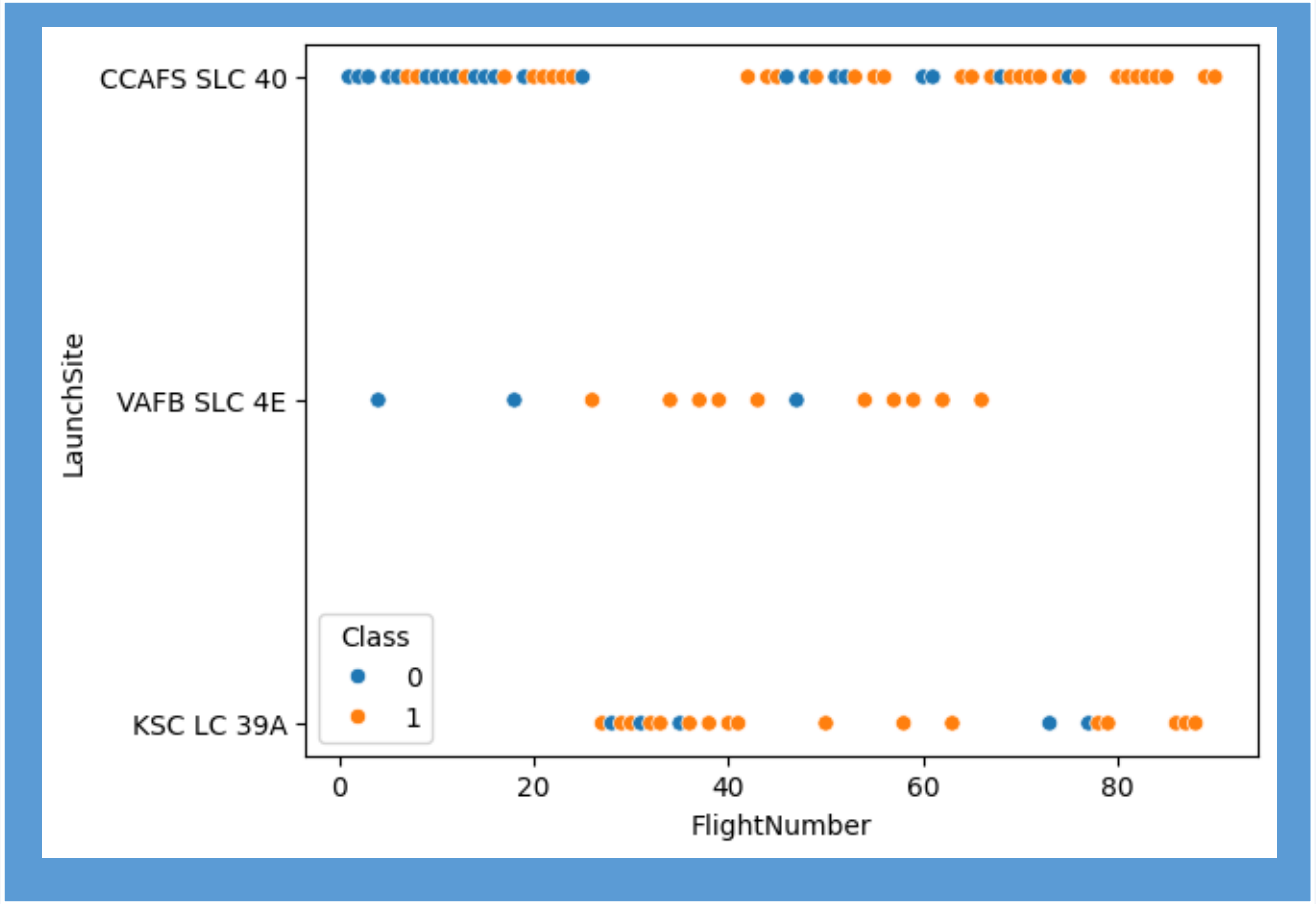
Results

- Predictive analysis results

Model	Training Accuracy	Test Accuracy
Logistic Regression	0.8333	0.8333
SVM	0.8857	0.8889
Decision Tree	0.8333	0.8333
K Nearest Neighbors	0.8333	0.8333


Section 2
Insights drawn from EDA

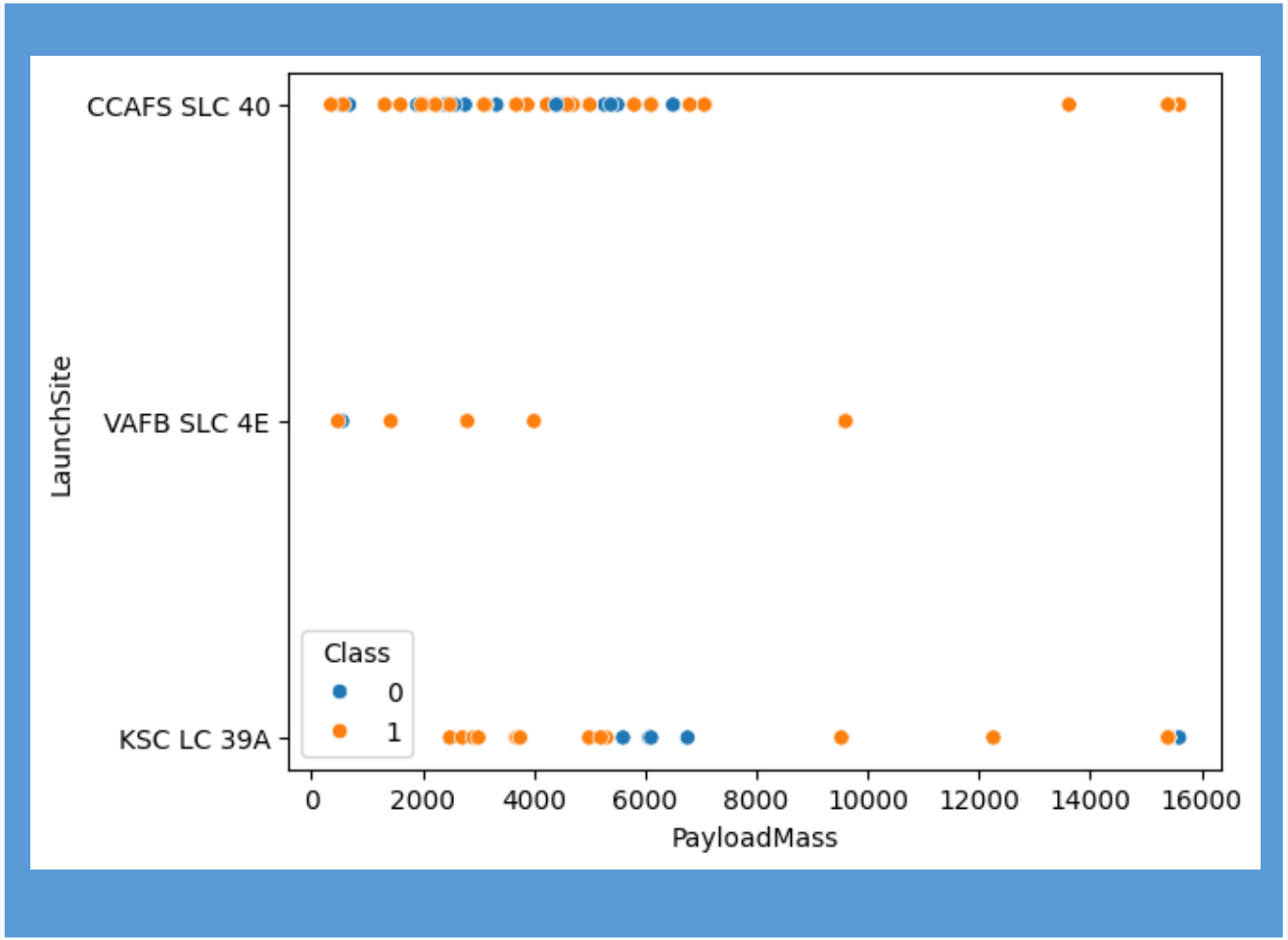
Flight Number vs. Launch Site



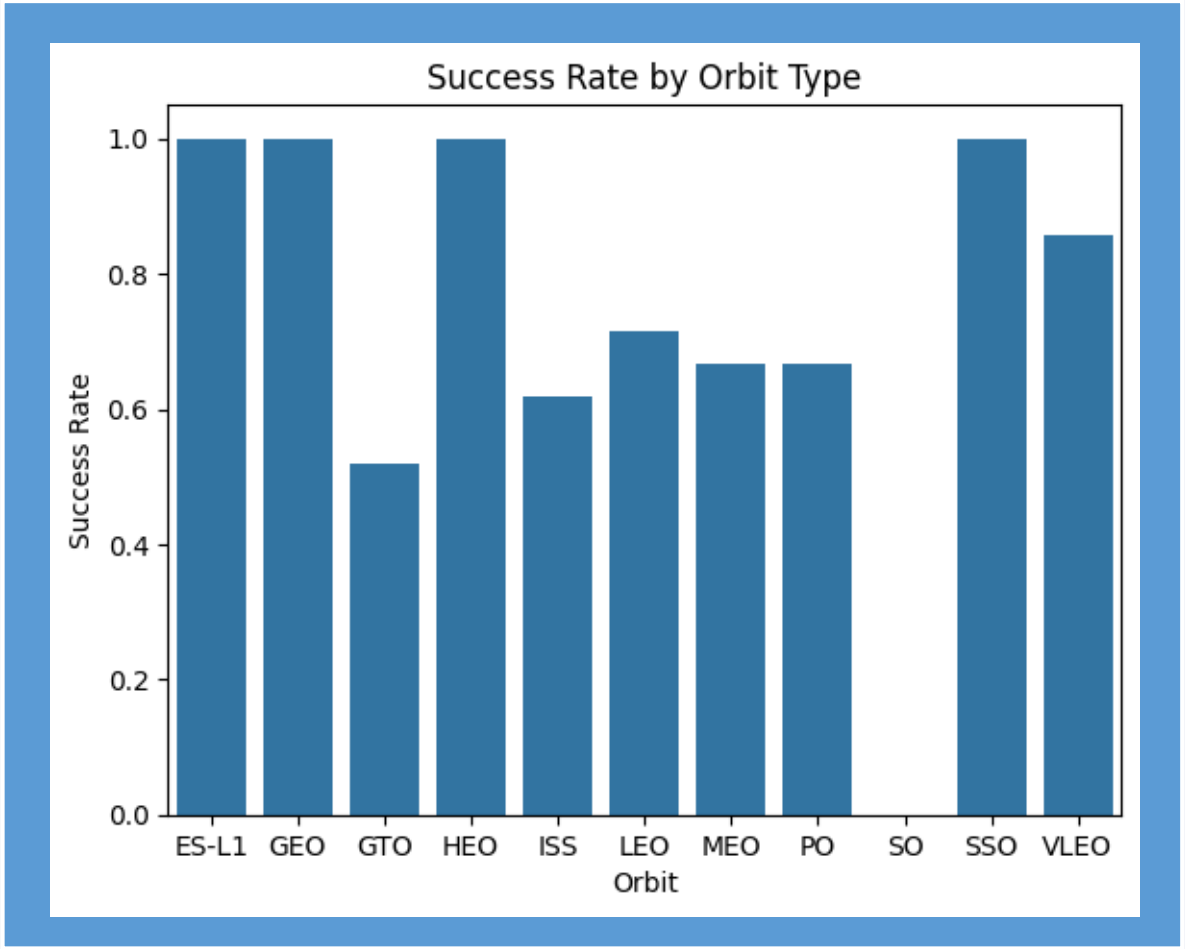
Findings: *The plot indicates that launch sites CCAFS SLC 40 and KSC LC 39A have higher success rates as the flight number increases, reflecting improvements in operations and technology at these sites. VAFB SLC 4E shows fewer launches but maintains a consistent success rate, possibly due to its specialization in specific mission types.*

Payload vs. Launch Site

Findings: *The scatter plot reveals that VAFB SLC 4E is primarily used for lighter payloads, with no launches exceeding 10,000 kg. In contrast, CCAFS SLC 40 and KSC LC 39A handle a wider range of payload masses, including heavier ones. This suggests that launch site selection is influenced by payload requirements and mission objectives.*



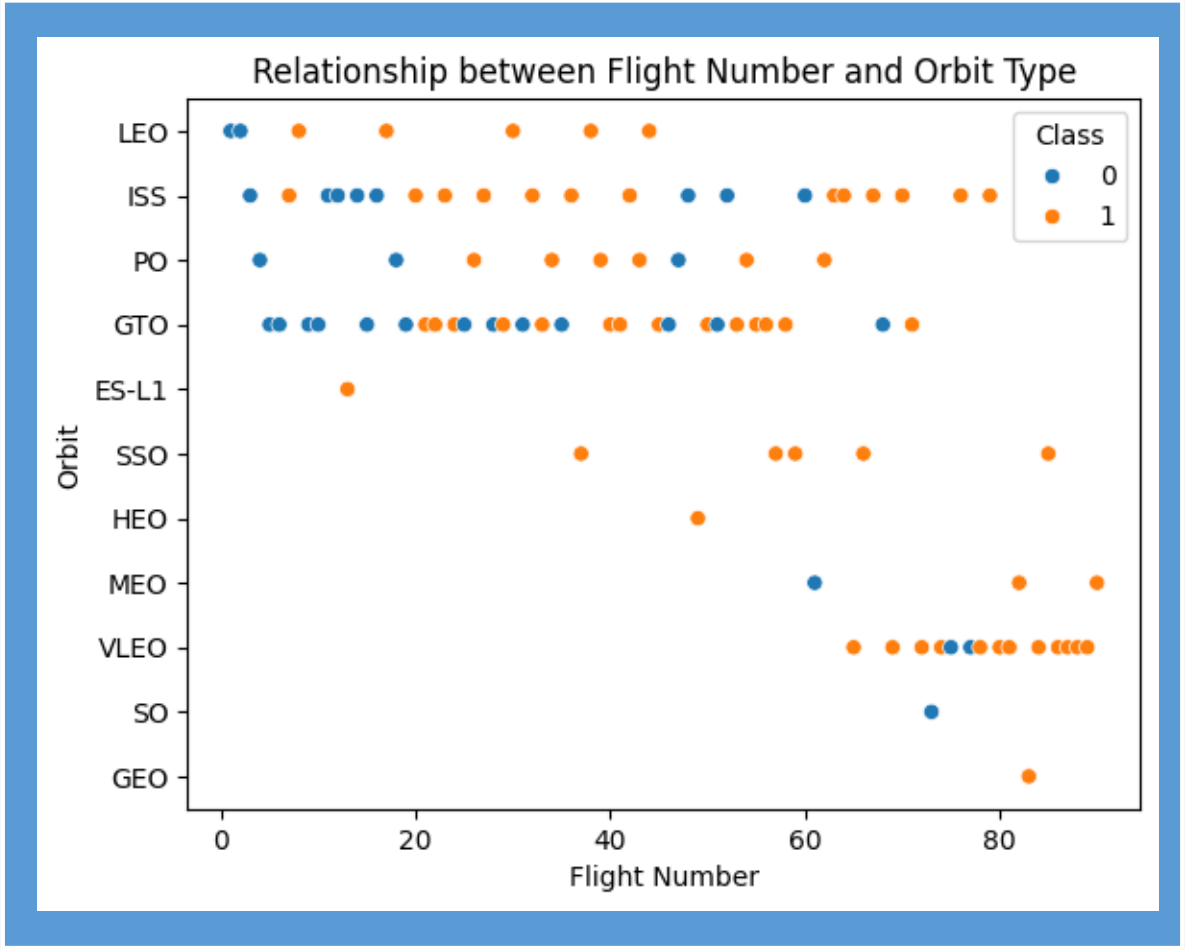
Success Rate vs. Orbit Type



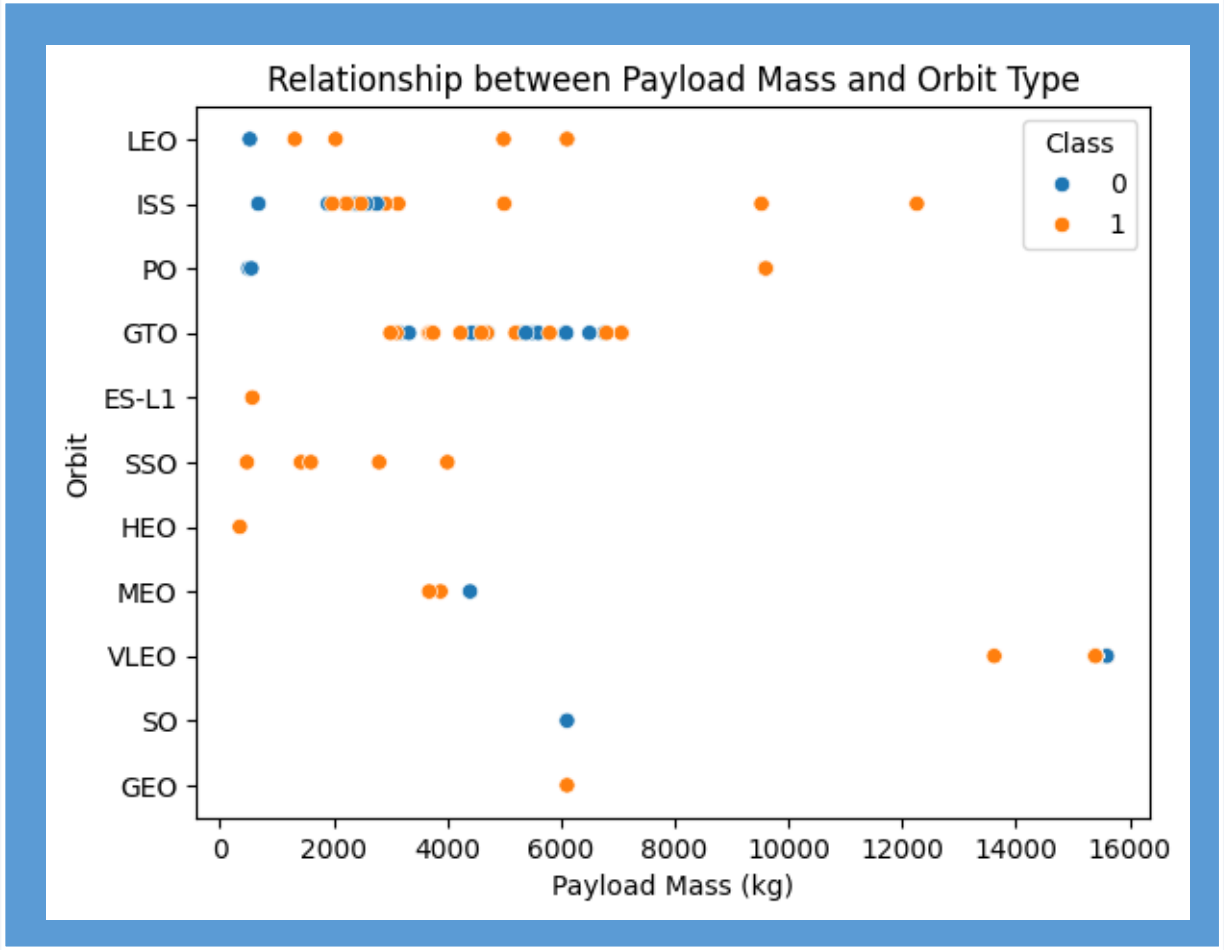
Findings: *The bar chart highlights that LEO (Low Earth Orbit) and ISS (International Space Station) orbits have the highest success rates, likely due to their lower complexity and shorter distances. GTO (Geostationary Transfer Orbit) shows more variability, indicating that higher-altitude orbits present greater challenges for successful landings.*

Flight Number vs. Orbit Type

Findings: *The scatter plot demonstrates that success rates in LEO orbits improve with increasing flight numbers, suggesting that experience and technological advancements have a positive impact. However, for GTO orbits, success rates do not show a clear correlation with flight number, indicating that other factors, such as payload mass or mission complexity, play a more significant role.*



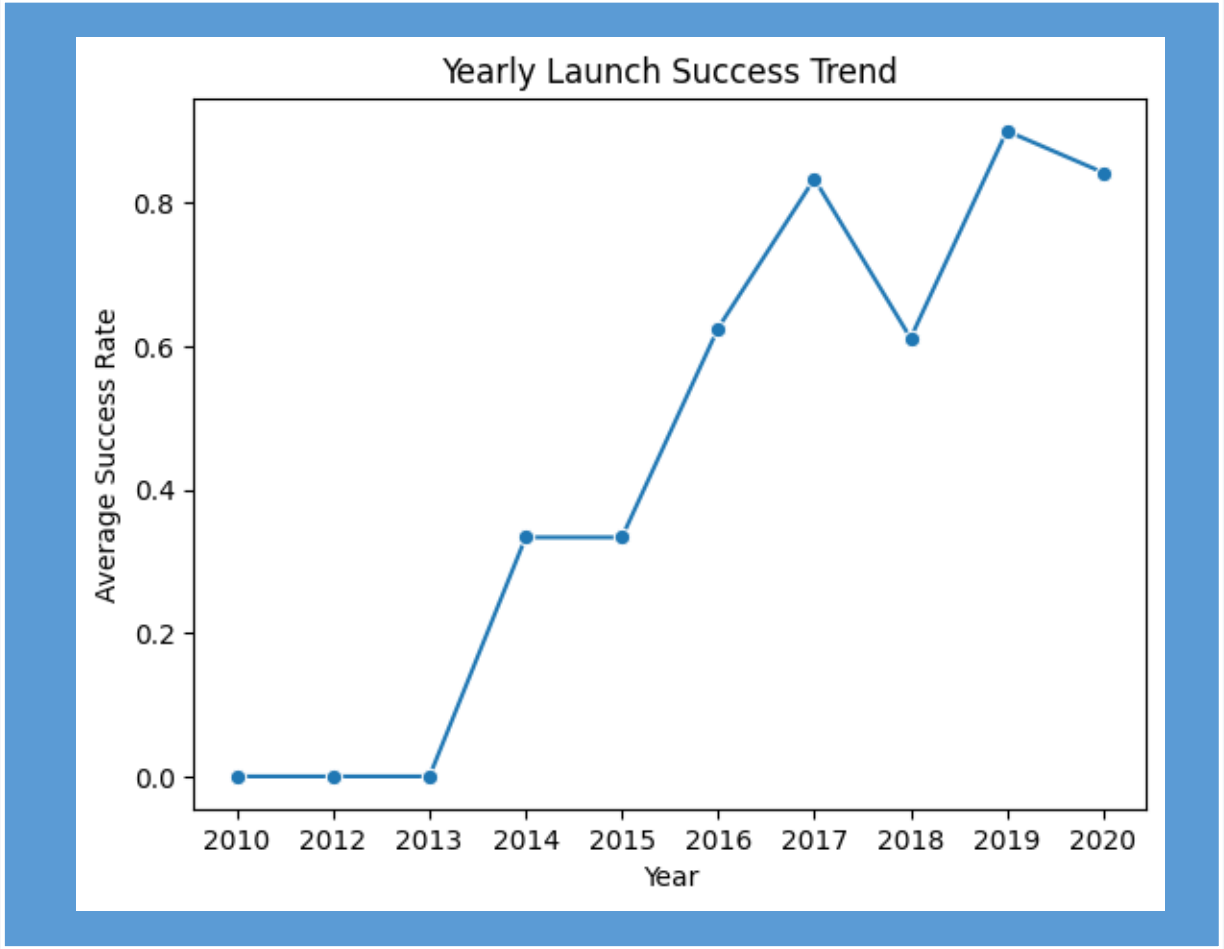
Payload vs. Orbit Type



Findings: *The scatter plot indicates that heavy payloads in Polar, LEO, and ISS orbits have higher success rates, possibly due to optimized launch profiles and lower mission complexity. For GTO orbits, both successful and unsuccessful landings are observed across a range of payload masses, suggesting that payload mass alone is not the sole determinant of success in these missions.*

Launch Success Yearly Trend

Findings: *The line chart shows a steady increase in success rates from 2013 onwards, with a notable improvement after 2015. This trend reflects SpaceX's continuous advancements in rocket technology and landing procedures, leading to more reliable and successful missions over time.*



All Launch Site Names

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

Out[12]:

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

Interpretation:

- The unique launch sites are **CCAFS LC-40**, **VAFB SLC-4E**, **KSC LC-39A**, and **CCAFS SLC-40**.
- This shows that SpaceX operates from multiple launch sites, each potentially serving different mission requirements.

Launch Site Names Begin with 'CCA'

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

Out[13]:

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

Interpretation:

- The first 5 records from launch sites starting with 'CCA' (CCAFS LC-40) were retrieved.
- These records include early missions, such as the Dragon Spacecraft Qualification Unit and NASA's COTS missions, showing SpaceX's initial focus on testing and collaboration with NASA.

Total Payload Mass

```
 * sqlite:///my_data1.db
Done.
Out[14]: Total_Payload_Mass
---------------------------
                     45596
```

Interpretation:

- The total payload mass carried for NASA (CRS) missions is **45,596 kg**.
- This indicates a significant contribution to NASA's cargo resupply missions to the ISS.

Average Payload Mass by F9 v1.1

```
 * sqlite:///my_data1.db
Done.
Out[15]: Average_Payload_Mass
---------------------------
                     2928.4
```

Interpretation:

- The average payload mass for the F9 v1.1 booster is **2,928.4 kg**.
- This suggests that the F9 v1.1 was used for medium-sized payloads, likely during the early stages of SpaceX's operational missions.

First Successful Ground Landing Date

```
 * sqlite:///my_data1.db
Done.
Out[16]: First_Successful_Landing_Date
---------------------------------------
                              2015-12-22
```

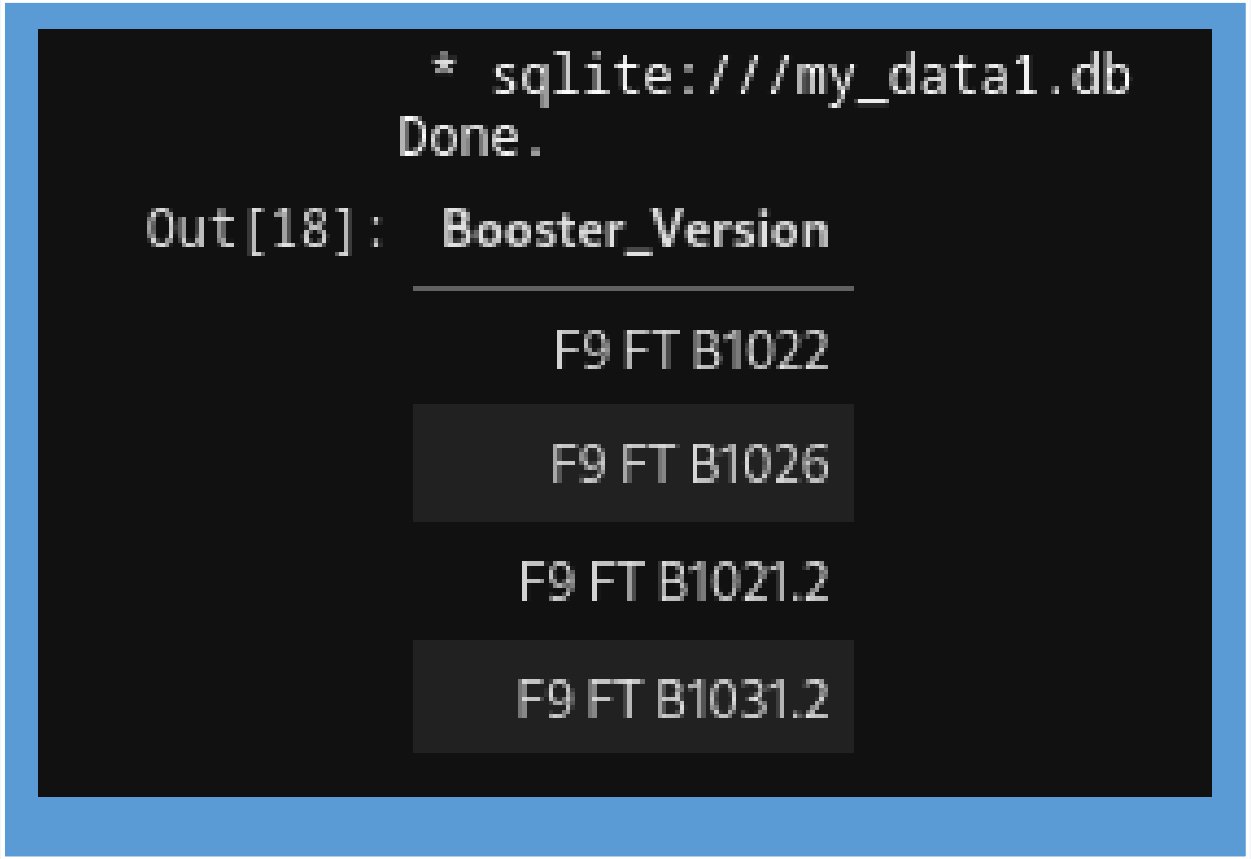
Interpretation:

- The first successful landing on a ground pad occurred on **2015-12-22**.
- This marks a significant milestone in SpaceX's ability to recover and reuse boosters, reducing launch costs.

Successful Drone Ship Landing with Payload between 4000 and 6000

Interpretation:

- The boosters **F9 FT B1022**, **F9 FT B1026**, **F9 FT B1021.2**, and **F9 FT B1031.2** successfully landed on drone ships with payloads in this range.
- This indicates that SpaceX has successfully managed to recover boosters carrying medium to heavy payloads, which is critical for cost-effective operations.



Total Number of Successful and Failure Mission Outcomes

```
 * sqlite:///my_data1.db
Done.
Out[19]:  Total_Successful_Missions  Total_Failed_Missions
                                98                     0
```

Interpretation:

- There were **98 successful missions** and **0 failed missions**.
- This highlights SpaceX's high success rate, which is a key factor in its competitive advantage.

Boosters Carried Maximum Payload

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

Out[20]:

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

Interpretation:

- The booster versions **F9 B5 B1048.4**, **F9 B5 B1049.4**, **F9 B5 B1051.3**, and others carried the maximum payload mass.
- These boosters are part of the Falcon 9 Block 5 series, which is designed for heavier payloads and reusability.

2015 Launch Records

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

Out[21]:

Month	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

Interpretation:

- In **January 2015**, the booster **F9 v1.1 B1012** failed to land on a drone ship from **CCAFS LC-40**.
- In **April 2015**, the booster **F9 v1.1 B1015** also failed to land on a drone ship from **CCAFS LC-40**.
- These failures occurred during the early stages of SpaceX's attempts to land boosters on drone ships, reflecting the challenges of this technology.

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

Interpretation:

- The most common outcome was **"No attempt"** (10 times), followed by **"Success (drone ship)"** and **"Failure (drone ship)"** (5 times each).
- This shows that during this period, SpaceX was still experimenting with landing techniques, with a mix of successes and failures.

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

Out[22]:

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

Section 3

Launch Sites Proximities Analysis

About launch sites

Key Findings:

- **Florida concentration:** *Most launch sites are located in Florida, specifically at **Cape Canaveral (CCAFS)** and **Kennedy Space Center (KSC)**. This is due to the proximity to the equator, which allows for more efficient launches into specific orbits.*
- **California Site:** *The **VAFB SLC-4E** site in California is primarily used for polar orbit missions, which complements SpaceX's launch capabilities.*

Brief Conclusion: *The map shows that SpaceX's launch sites are strategically located to optimise missions according to desired orbits. The concentration in Florida allows efficient launches into equatorial orbits, while the site in California facilitates polar missions. This geographic distribution is key to SpaceX's operational flexibility and efficiency.*

Locations and access points

Key Findings:

1. Launch Markers:

- Green Markers: Indicate successful launches.
- Red Markers: Indicate failed launches.
- Yellow Marker: Indicates a launch in progress or a different status (not specified).

2. Clusters:

- The orange circles represent clusters of markers. These clusters group multiple launches in the same area for easier visualization.

3. Launch Sites:

- There are two main launch sites marked on the map.
- The upper site has a cluster with 4 launches (3 successful and 1 failed).
- The lower site has a cluster with 3 launches (2 successful and 1 in progress or a different status).

4. Locations and Access Points:

- The map shows roads, parking areas, and buildings near the launch sites.
- There is a building labeled "LOOPS BLDG 40" near the upper launch site.

Brief Conclusion: The map displays two main launch sites with clusters indicating the success or failure of launches. The upper site has a higher number of launches with a high success rate (3 out of 4), while the lower site has fewer launches but also shows mostly successful outcomes (2 out of 3, with one in progress or a different status). The visualization allows for quick identification of areas with better performance and potential patterns related to launch outcomes.

Around launch sites

Key Findings:

1. Mouse Position (MousePosition):

- The `MousePosition` tool allows you to interactively obtain specific coordinates on the map. This is useful for measuring distances between points of interest.

2. Markers:

- Markers indicate specific points on the map, such as launch sites and proximity points.

3. PolyLines:

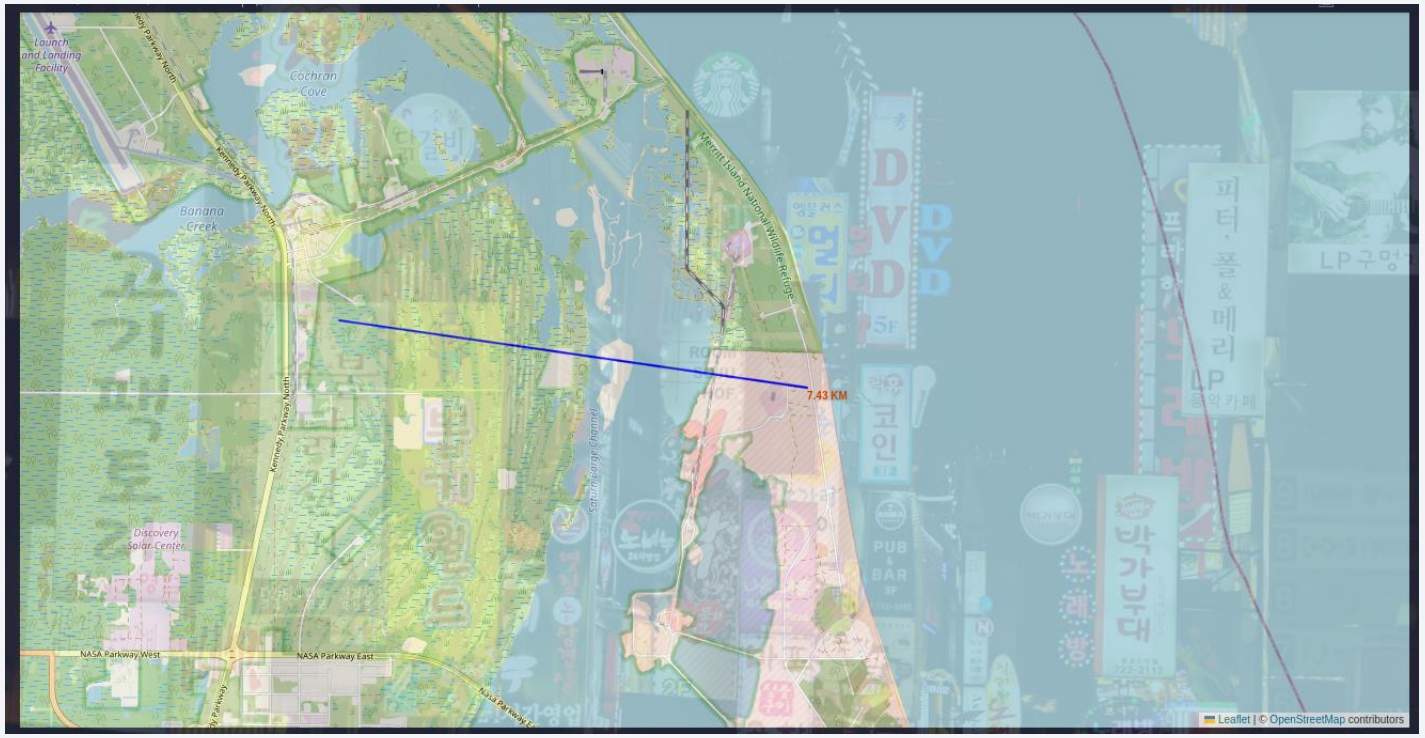
- `PolyLines` visually connect the launch site to selected proximity points. This helps in visualizing and measuring the distances between these points.

4. Measured Distance:

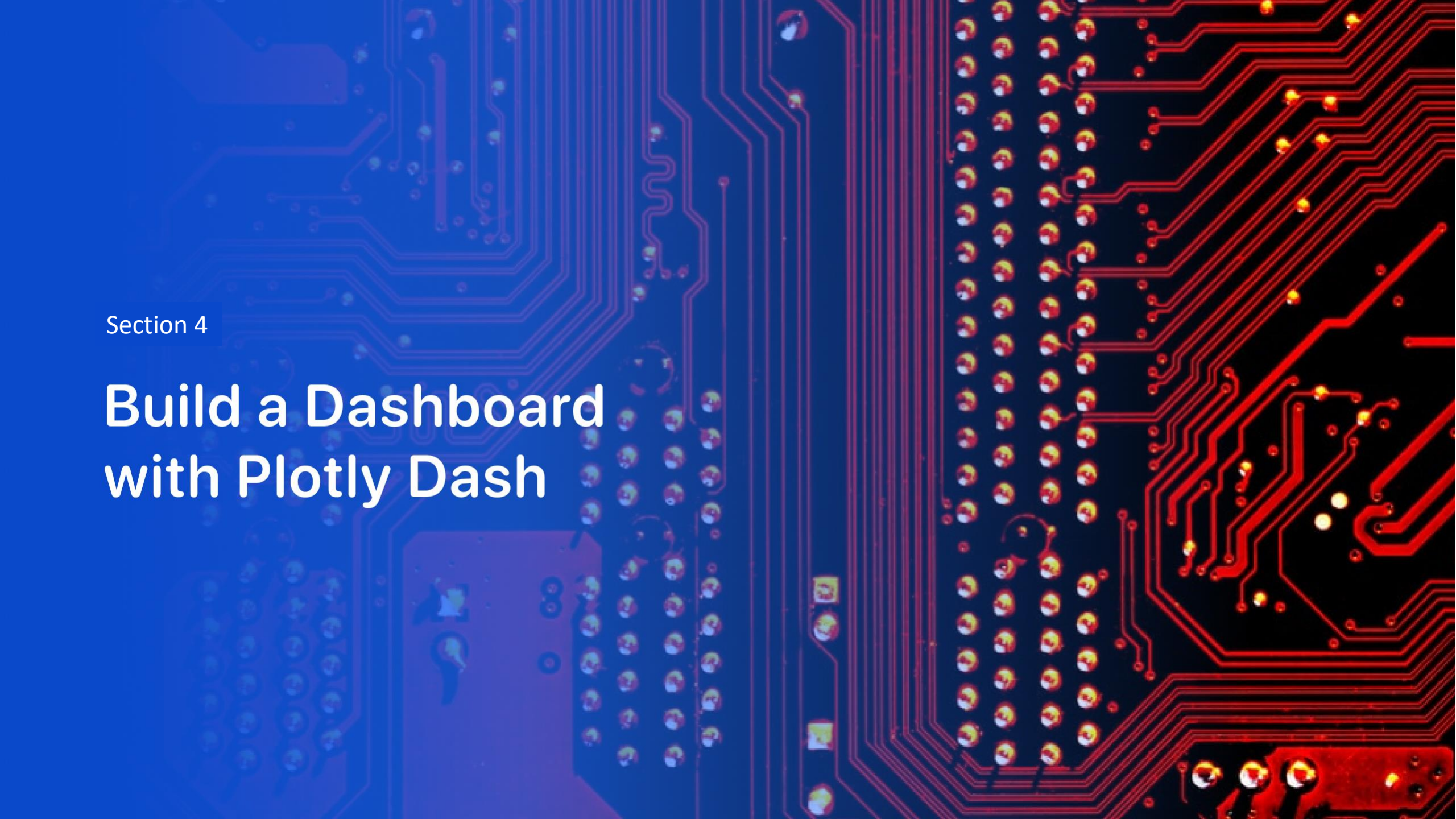
- The map shows a blue line connecting a launch site to a proximity point, with a measured distance of 7.43 km.

5. Locations and Access Points:

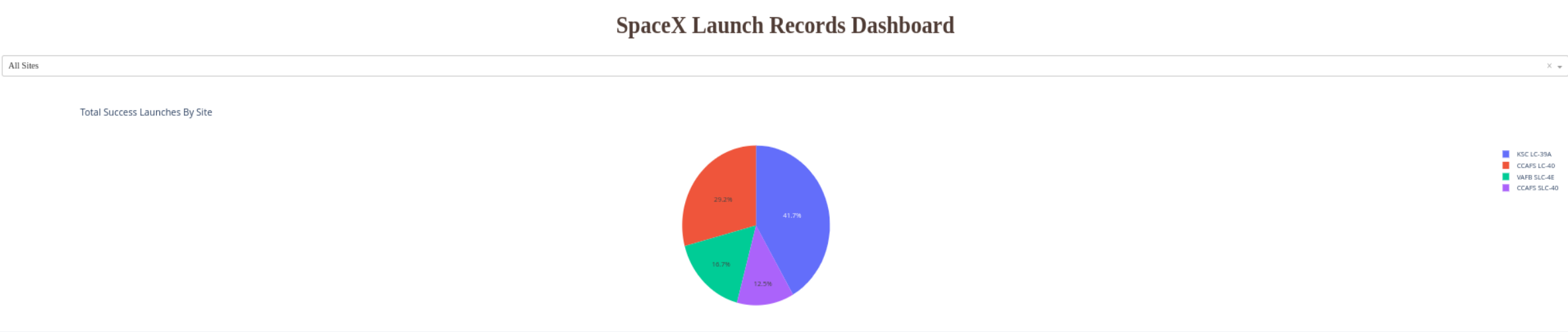
- The map displays various locations and access points, including roads, buildings, and natural areas. This provides context about the environment around the launch site.



Brief Conclusion: The map shows a launch site connected to a proximity point with a blue line, indicating a distance of 7.43 km. The `MousePosition` tool allows you to obtain specific coordinates interactively, while markers and `PolyLines` help visualize and measure distances between the launch site and its proximities. This analysis is useful for evaluating proximity factors that may affect the launch success rate, providing a more detailed understanding of the environment and its potential impacts on launch operations.


Section 4
Build a Dashboard with Plotly Dash

All launch sites findings



Launch Site Distribution:

- **KSC LC-39A (41.7%):** *The Kennedy Space Center Launch Complex 39A is the first most used site. Its significant usage suggests that it is another critical launch pad for SpaceX, possibly due to its historical significance (used for Apollo and Space Shuttle missions) and modern upgrades to support SpaceX's missions.*
- **CCAFS SLC-40 (29.2%):** *This site, located at Cape Canaveral Air Force Station, indicates that it has been a primary and reliable launch site for SpaceX. The high percentage could be due to its strategic location, established infrastructure, and favorable launch conditions.*
- **VAFB SLC-4E (12.5%):** *Vandenberg Air Force Base Space Launch Complex 4E has a smaller share of launches. This site is typically used for launches that require polar orbits, which are less frequent but essential for specific missions like satellite deployments.*
- **Boca Chica (10.7%):** *Boca Chica, the newest and least used site, is where SpaceX's Starship development and testing occur. The low percentage is expected as Starship is still in its development and testing phases.*

Future Implications:

- **Expansion and Innovation:** *The presence of Boca Chica, despite its low percentage, highlights SpaceX's commitment to innovation and future space exploration. As Starship development progresses, we can expect an increase in launches from this site.*
- **Operational Flexibility:** *The ability to launch from multiple sites provides operational flexibility and resilience. This is crucial for maintaining a high launch cadence and meeting the growing demand for space services.*
- **Geographical Advantages:** *Each launch site offers unique geographical advantages. For example, Cape Canaveral and Kennedy Space Center are ideal for equatorial orbits, while Vandenberg is suitable for polar orbits. This geographical diversity allows SpaceX to cater to a wide range of mission profiles.*

Highest succes launch site KSC LC-39A

1. Success vs. Failure Rate:

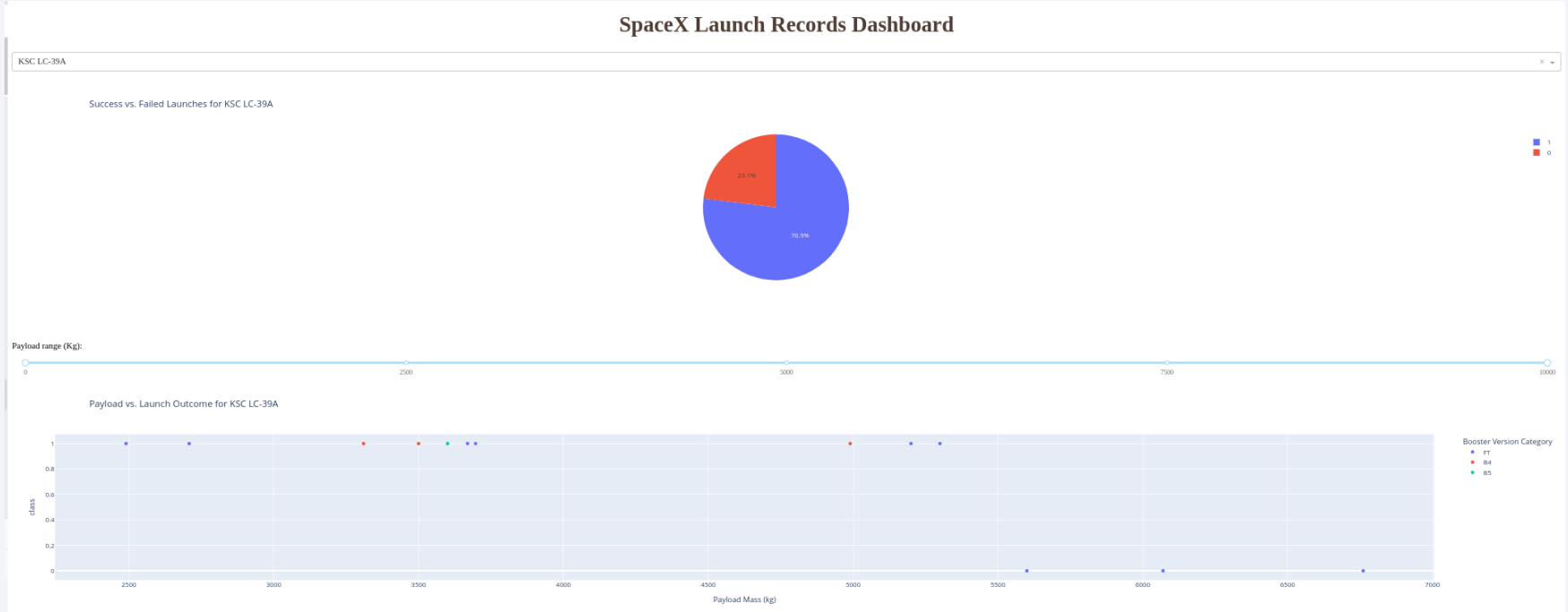
- *The majority of launches from KSC LC-39A have been successful, with a success rate of 76.9%. This indicates high reliability and efficiency in launch operations at this site.*
- *The failure rate of 25.1% is relatively low but still significant and warrants a more detailed analysis to identify the causes of failures and improve the success rate.*

2. Relationship Between Payload Mass and Launch Outcome:

- *The scatter plot shows that most successful launches (blue dots) are distributed across the entire range of payload masses, from light to heavy loads.*
- *Failed launches (red dots) appear to be more scattered and less frequent, but there is no clear correlation between payload mass and launch outcome. This suggests that factors other than payload mass may be contributing to the failures.*

3. Operational Implications:

- *The high success rate at KSC LC-39A reinforces confidence in this launch site for future missions.*
- *Analyzing the failures can help identify areas for improvement in launch procedures, the technology used, or environmental conditions that may affect launch success.*
- *The diversity in payload mass indicates that KSC LC-39A is versatile and can handle a variety of missions, from small satellites to heavy payloads.*



Efficiency in payload over launch outcome

1. Distribution of Successful and Failed Launches:

- Most of the points on the graph are blue, indicating that the majority of launches have been successful.

- The red dots, representing failed launches, are less frequent and scattered across the payload mass range.

2. Relationship Between Payload Mass and Launch Outcome:

- There is no clear correlation between payload mass and launch outcome. Both successful and failed launches occur across the entire range of payload masses.

- This suggests that payload mass is not the primary determining factor for the success or failure of a launch. Other factors, such as environmental conditions, technology used, and launch procedures, may have a more significant impact.

3. Performance of Different Rocket Versions:

- The graph shows that different rocket versions (F9, F9+, FH) have been used for a variety of payload masses.

- Most successful launches appear to be associated with the Falcon 9 and Falcon 9 Full Thrust versions, indicating their reliability and efficiency.

- The Falcon Heavy (FH) also has successful launches, but its use seems less frequent compared to the Falcon 9 versions.

4. Operational Implications:

- The overall high success rate reinforces confidence in SpaceX's launch operations.

- Analyzing the failures can help identify areas for improvement in launch procedures, the technology used, or environmental conditions that may affect launch success.

- The diversification in payload mass and the use of different rocket versions indicate that SpaceX is versatile and can handle a variety of missions, from small satellites to heavy payloads.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

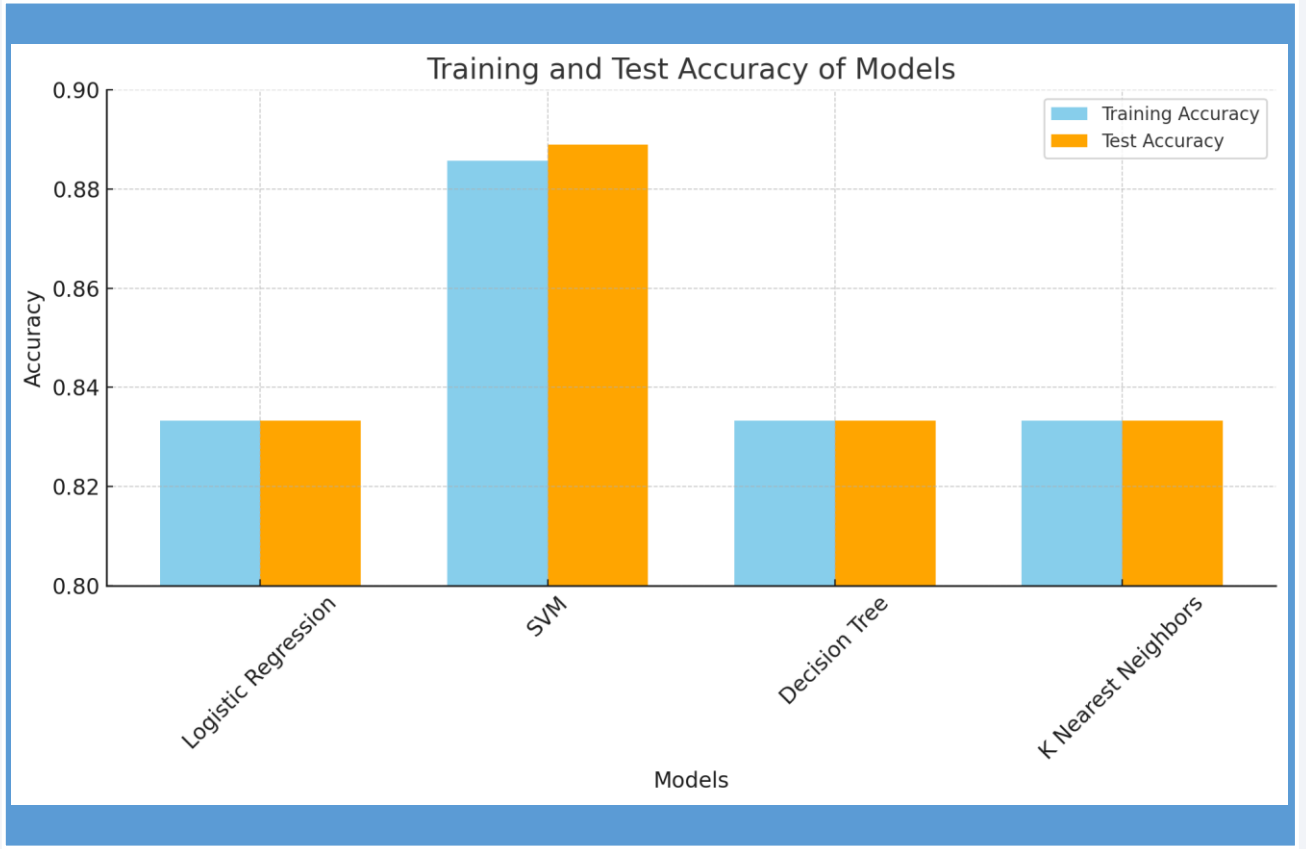
1. Logistic Regression:
- Observation: The training and test accuracies are nearly identical, indicating that the model generalizes well to unseen data without overfitting.

2. SVM (Support Vector Machine):
- Observation: SVM shows the highest accuracy among the models, with both training and test accuracies being very close. This suggests that SVM is the most effective model for this dataset, balancing well between training and generalization.

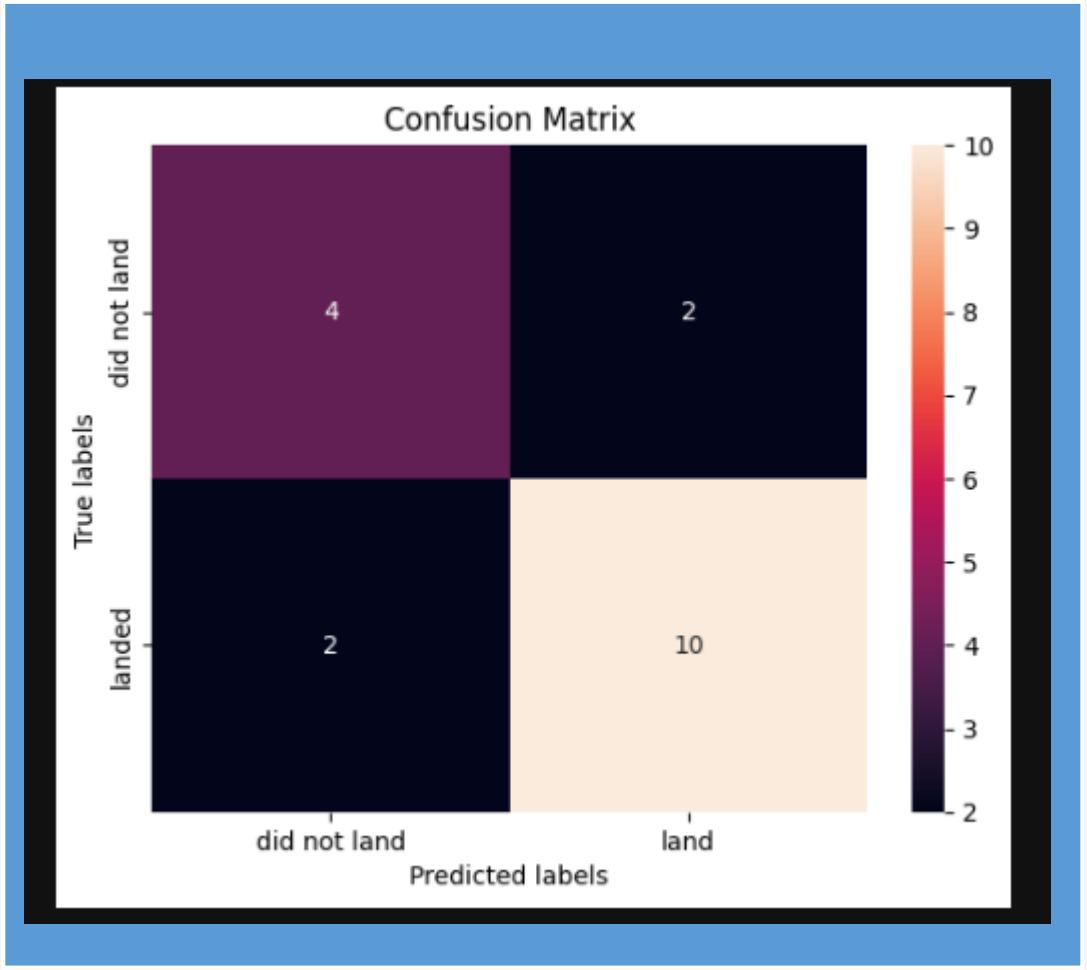
3. Decision Tree:
- Observation: Similar to Logistic Regression, the Decision Tree model has nearly identical training and test accuracies, indicating good generalization. However, its accuracy is lower compared to SVM.

4. K Nearest Neighbors:
- Observation: K Nearest Neighbors also shows similar training and test accuracies, suggesting good generalization. Its performance is on par with Logistic Regression and Decision Tree but lower than SVM.

Conclusion:
- **Best Performing Model:** *SVM is the best performing model with the highest training and test accuracies, indicating it is the most suitable model for this dataset.*
- **Generalization:** *All models show good generalization as the training and test accuracies are very close to each other.*



Confusion Matrix



Confusion Matrix Analysis in Business Terms

1. True Positives (TP) - 10:
- **Meaning:** *The model correctly predicted that 10 shipments would reach their destination.*
- **Impact:** *This is positive because the company can confidently plan for the delivery of these shipments, ensuring that resources and staff are available at the right time.*

2. True Negatives (TN) - 4:
- **Meaning:** *The model correctly predicted that 4 shipments would not reach their destination.*
- **Impact:** *This allows the company to take preventive measures, such as rescheduling shipments or notifying customers about potential delays, improving expectation management and reducing customer dissatisfaction.*

3. False Positives (FP) - 2:
- **Meaning:** *The model incorrectly predicted that 2 shipments would reach their destination when they actually did not.*
- **Impact:** *This can lead to poor planning and customer dissatisfaction, as delivery was expected but did not occur. The company may need to manage these cases with clear communications and possible compensations.*

4. False Negatives (FN) - 2:
- **Meaning:** *The model incorrectly predicted that 2 shipments would not reach their destination when they actually did.*
- **Impact:** *Although this may seem less problematic, it can lead to underutilization of resources and missed opportunities to optimize logistics. The company could have planned better if it had known these shipments would arrive.*

Conclusions

In business terms, the SVM model shows good performance with high precision and recall, which is crucial for logistical planning and customer satisfaction. However, false positives and negatives indicate areas where the company may need to better manage customer expectations and optimize resources. Improving the model's overall accuracy can lead to greater operational efficiency and a better customer experience.

Importance

- **Efficient Planning:** Accurate predictions enable better planning of resources and staff.
- **Customer Satisfaction:** Correctly identifying successful shipments and potential delays improves expectation management and reduces customer dissatisfaction.
- **Resource Optimization:** Minimizing false negatives can lead to better resource utilization and more efficient logistics.
- **Informed Decision-Making:** Evaluation metrics provide valuable insights for making informed decisions about model improvement and process optimization.

Appendix

These slides correspond to the final project for the:

‘IBM Data Science Professional Certificate’

for more information see this **GitHub repository**

Thank you!