



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hsueh-Lin Yu
2025/01/13

Outline

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

GitHub repository: <https://github.com/lennexyu/IBM-Module-10>

Executive Summary

- This research investigates the factors contributing to a successful rocket landing through comprehensive data analysis and predictive modeling. The study employs methodologies to collect, process, and analyze data from SpaceX, with insights aimed at enhancing rocket landing success.
- Summary of methodologies
 - Data Collection: Utilized SpaceX REST API and web scraping techniques to compile relevant data.
 - Data Wrangling: Created a success/failure outcome variable to streamline analysis.
 - Exploratory Analysis: Employed visualization techniques to explore relationships between payload, launch site, flight number, and yearly trends.
 - Statistical Analysis: Leveraged SQL to calculate key metrics, including total payload, payload ranges for successful launches, and success/failure rates.
- Geographical Insights: Analyzed launch site success rates relative to proximity to coasts and other geographical markers.
- Predictive Modeling: Built logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN) models to predict landing outcomes.
- Summary of all results
 - Launch Success Trends: Success rates have improved significantly over time.
 - Top Performing Launch Site: KSC LC-39A exhibits the highest success rate among all landing sites.
 - Orbit Performance: Orbits such as ES-L1, GEO, HEO, and SSO maintain a 100% success rate.
 - Geographical Patterns: Most launch sites are located near the equator


Introduction

Background

SpaceX, a pioneer in the space industry, aims to make space travel accessible and affordable for all. Its notable achievements include sending spacecraft to the International Space Station, deploying a satellite constellation to provide global internet access, and conducting manned space missions. SpaceX's ability to achieve these milestones at a lower cost (\$62 million per launch) is largely due to its innovative reuse of the Falcon 9 rocket's first stage. In contrast, other providers, unable to reuse their first stages, incur costs exceeding \$165 million per launch. Predicting whether the first stage will successfully land allows us to estimate the cost of a launch. By leveraging publicly available data and machine learning models, we can forecast the reusability of the first stage for SpaceX and its competitors, offering valuable insights into launch cost optimization.

Research Questions

- What factors influence the success of a Falcon 9 rocket's first-stage landing?
- Can machine learning models predict whether the Falcon 9's first stage will successfully land?



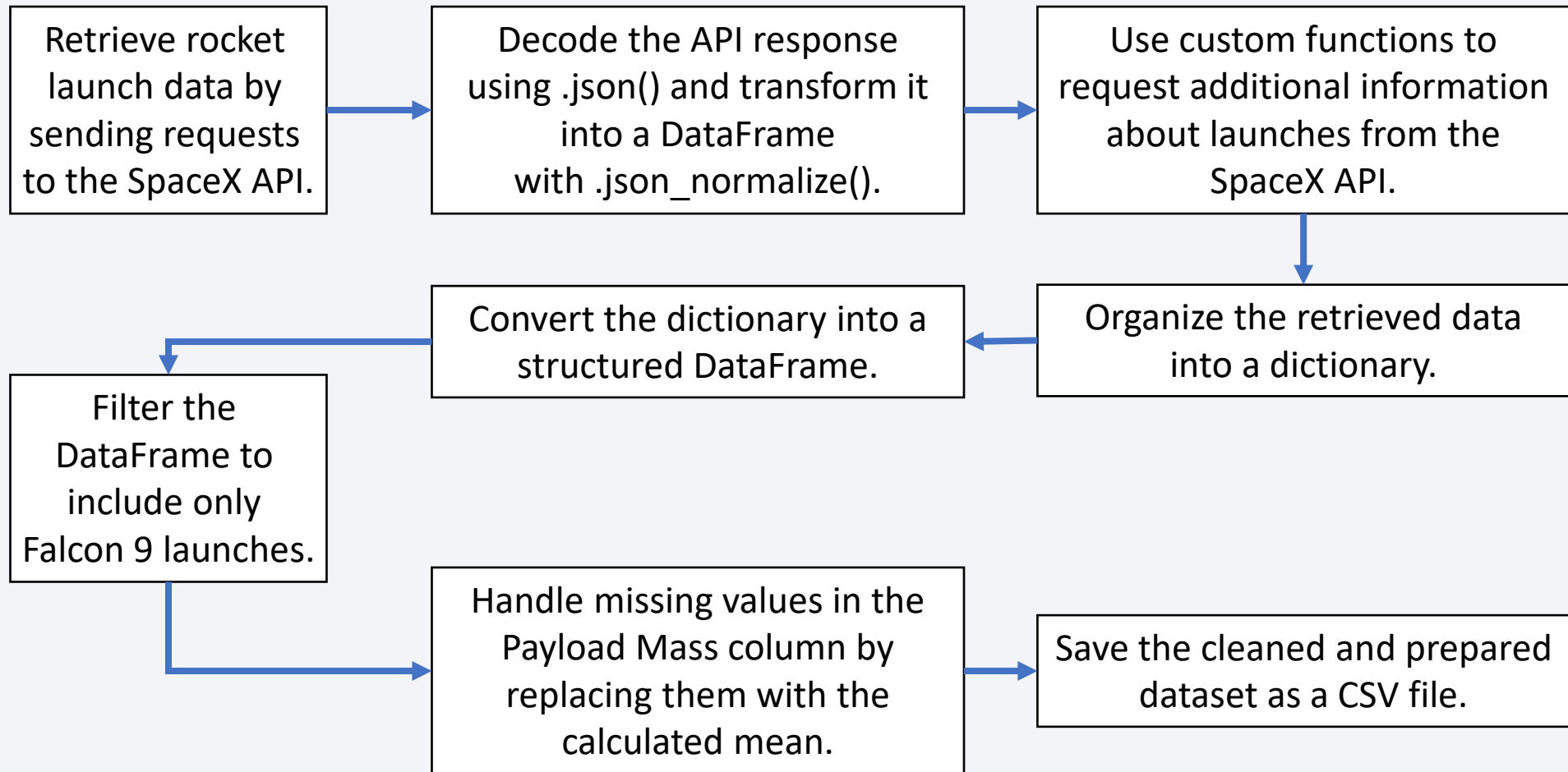
Section 1

Methodology

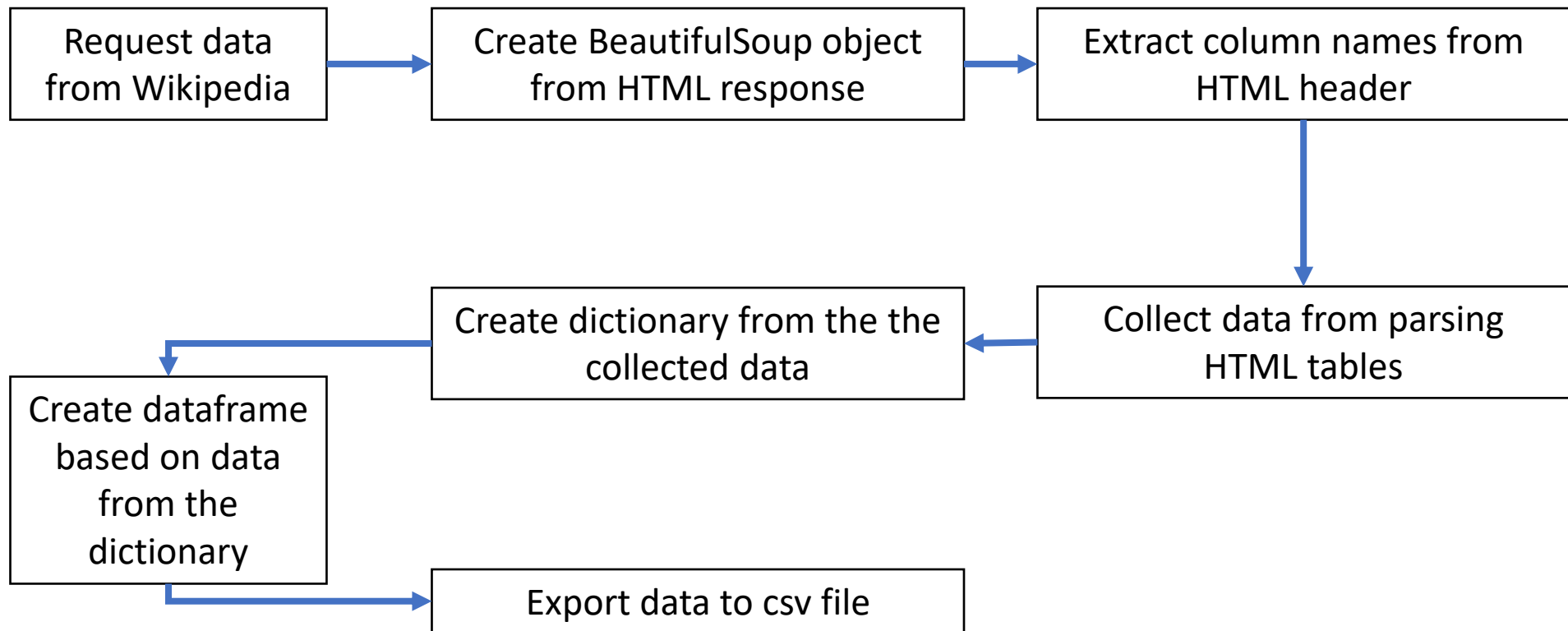
Methodology

- Collect data from the SpaceX REST API and supplement it with information obtained through web scraping.
- Prepare the data for analysis and modeling by performing data wrangling, which includes filtering relevant data, handling missing values, and applying one-hot encoding.
- Conduct exploratory data analysis (EDA) using SQL and various data visualization techniques to uncover patterns and insights.
- Create interactive visualizations with Folium and Plotly Dash to better understand and communicate the data.
- Develop classification models to predict landing outcomes, optimizing their performance through hyperparameter tuning and evaluation to identify the best-performing model.

Data Collection – SpaceX API



Data Collection - Scraping



Data Wrangling

1. Library Imports	pandas and numpy were imported for data manipulation and array operations
↓	
2. Data Loading	A CSV file was read into a DataFrame using <code>pd.read_csv()</code> .
↓	
3. Initial Exploration	Basic checks included displaying the first few rows with <code>df.head(10)</code> and calculating the percentage of missing values in each column with <code>df.isnull().sum()/len(df)*100</code> .
↓	
4. Value Counts	<code>value_counts()</code> was applied to the columns LaunchSite, Orbit, Outcome to understand the distribution of values in these columns.
↓	
5. Outcome Classification	Outcome Classification: A <code>bad_outcomes</code> set was created by selecting specific outcomes that correspond to unsuccessful landings. A new column, <code>landing_class</code> , was generated using <code>apply</code> with a lambda function: rows with outcomes in <code>bad_outcomes</code> =0, while other outcomes = 1.
↓	
6. Landing Class Summary	The mean of <code>landing_class</code> was calculated with <code>df["landing_class"].mean()</code> to summarize landing success rate.

EDA with Data Visualization

- **Flight Number vs. Launch Site:**
 - To understand the distribution of flight numbers across different launch sites and assess their success rates. This revealed that earlier flights (smaller flight numbers) were more prone to failure, and specific sites like CCAFS SLC 40 had the highest number of launches but the lowest success rate.
- **Payload vs. Launch Site**
 - To analyze the relationship between payload mass and success rates at different launch sites. It was observed that higher payloads (>12,000kg) at CCAFS SLC 40 and lower payloads (<5,000kg) at KSC LC 39A had higher success rates.
- **Success Rate vs. Orbit Type**
 - To evaluate success rates for different orbit types. Certain orbits (e.g., ES-L1, GEO, HEO, and SSO) showed 100% success rates, while others had varied outcomes.
- **Flight Number vs. Orbit Type**
 - To track the testing progression of orbit types over time. Early tests focused on LEO, ISS, PO, and GTO, with failures more common in GTO.
- **Payload vs. Orbit Type**
 - To investigate payload mass differences across orbit types and their association with success rates. For example, lower payload masses in LEO, ISS, and PO were linked with failures, while VLEO tested the highest payloads.
- **Launch Success Yearly Trend**
 - To study the temporal trend of success rates over the years. This showed an improving trend from 2013 to 2017, with setbacks in 2018, and an overall increase in success rates up to 2020.

EDA with SQL

- **Display:**

- Names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1

- **List:**

- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster_versions which have carried the maximum payload mass.
- Failed landing records with month, booster version and launch sites in year 2015.
- Count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

- GitHub URL: https://github.com/lennexyu/IBM-Module-10/blob/main/jupyter-labs-eda-sql-coursera_sqlite.ipynb

Build an Interactive Map with Folium

1. Blue Circle Marker for NASA Johnson Space Center:

- Added at NASA Johnson Space Center using its latitude and longitude coordinates.
- A popup label displays its name to provide clear identification for the site.
- To highlight NASA Johnson Space Center as a key location.

2. Red Circle Markers for Launch Sites:

- Added at all launch site coordinates, each with a popup label showing the launch site's name.
- To mark the locations of all launch sites and make them easily identifiable on the map.

3. Green and Red Markers for Successful and Unsuccessful Launches, respectively.

- To visually present the sites with higher success rate.

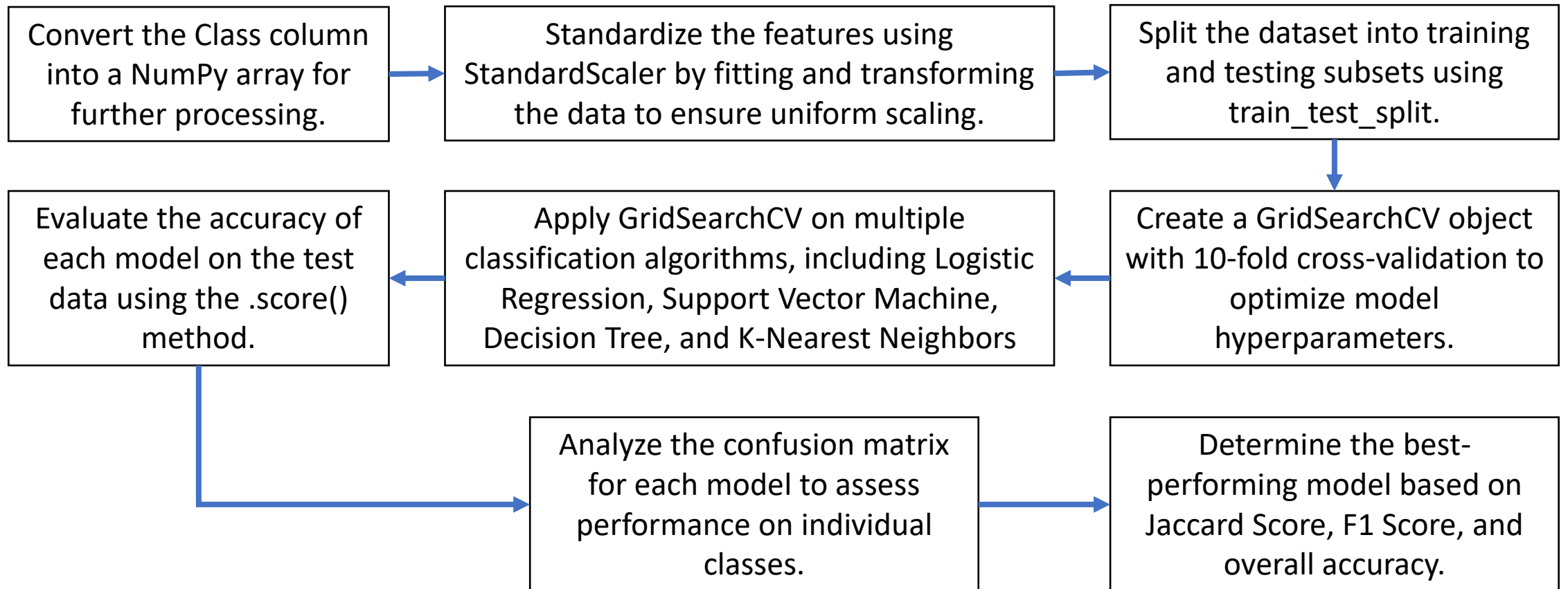
4. Colored Lines for Proximity Connections


- Connecting CCAFS SLC-40 to the nearest coastline, railway, highway, and city.

Build a Dashboard with Plotly Dash

- **Launch Site Selection by Dropdown List:** Enable users to choose between viewing data for all launch sites or filtering by a specific launch site.
- **Payload Mass Range Slider:** Provide an interactive slider for users to adjust and select the desired range of payload mass.
- **Pie Chart for Launch Outcomes:** Display the distribution of successful and unsuccessful launches as percentages of the total, allowing users to analyze overall performance.
- **Scatter Plot for Payload vs. Success by Booster Version:** Visualize the relationship between payload mass and launch success, categorized by booster version, to help users identify correlations.

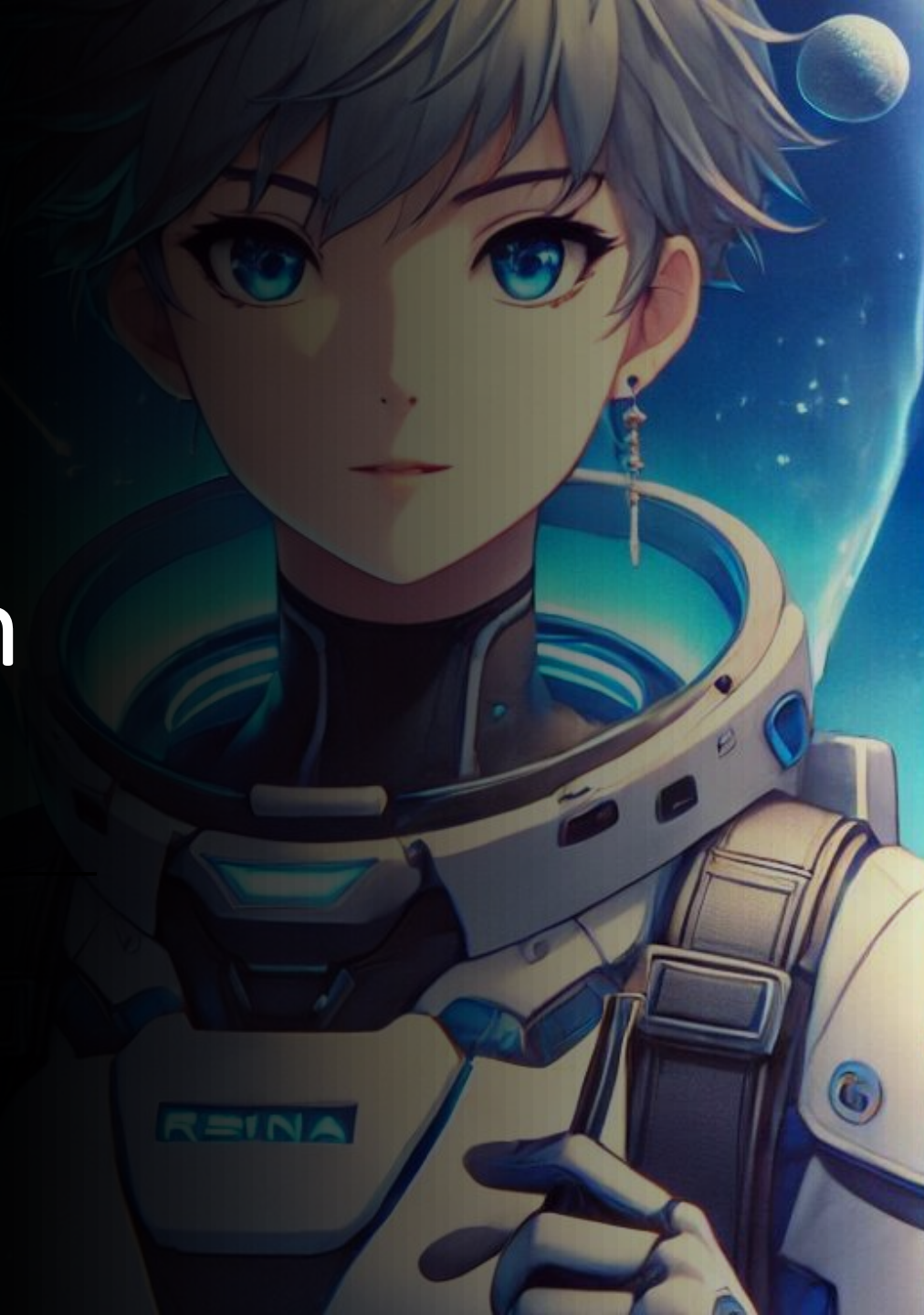
Predictive Analysis (Classification)





Section 2

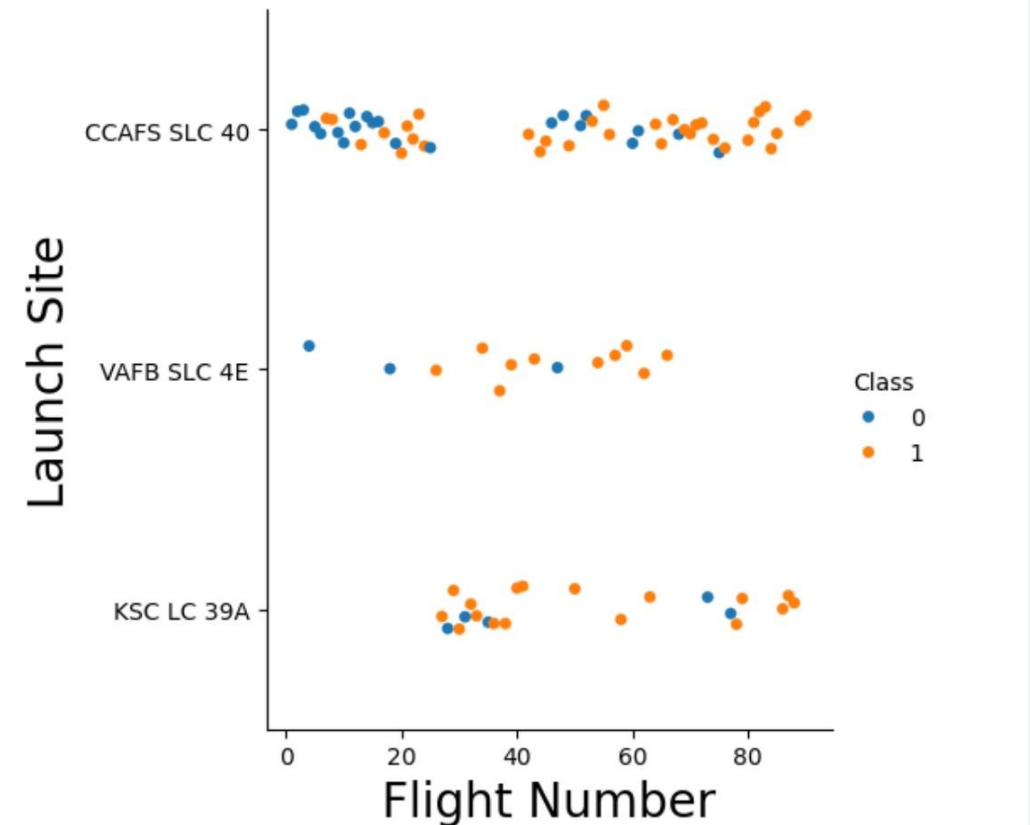
Insights drawn from EDA



Flight Number vs. Launch Site

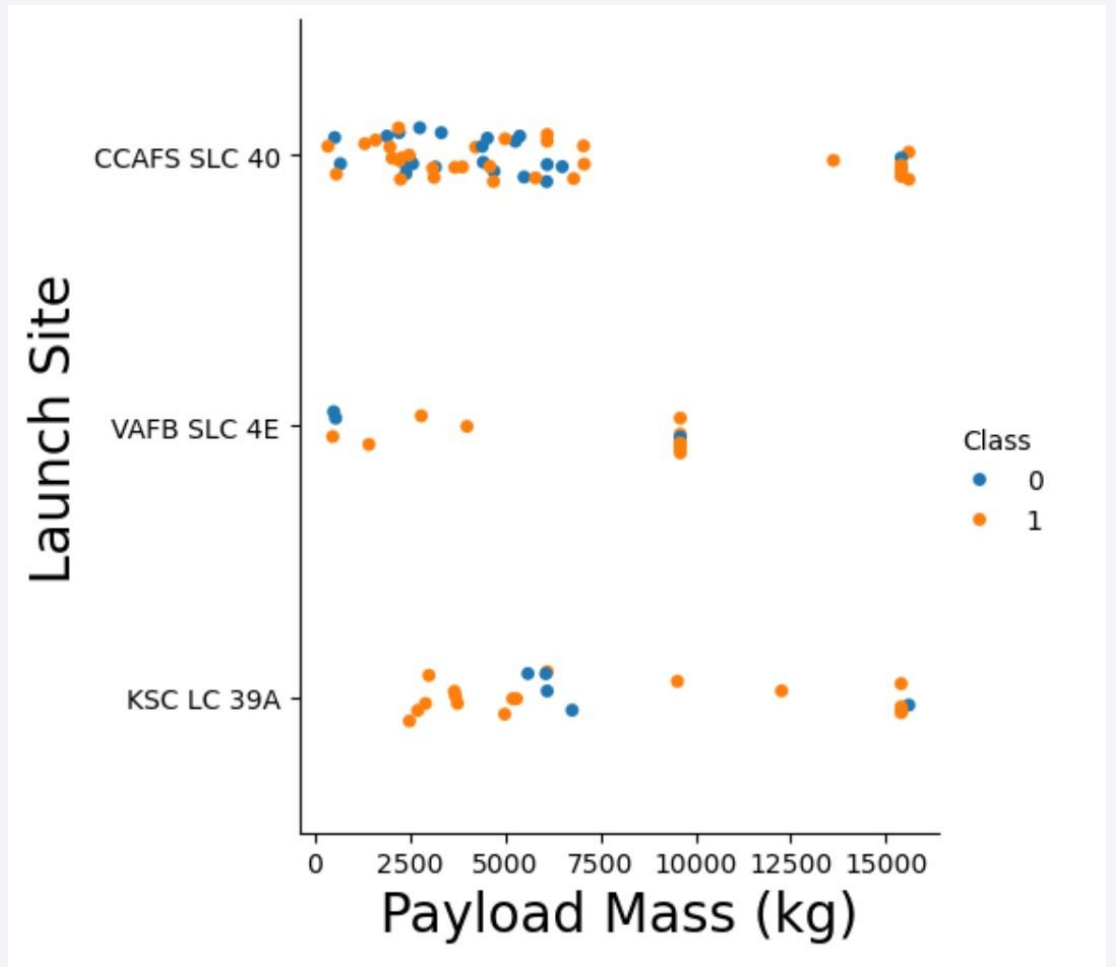
- Smaller flight number = earlier flights
- Larger flight number = later flights
- Failure = blue; Success = orange
- There were more failed events in earlier flights (blue color).
- Most flights occurred at CCAFS SLC 40, with lowest success rate (60%).
- Less flights occurred at VAFB SLC 4E and KSC LC 39A, with success rates of 76.92% and 77.27%, respectively.

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect=1)  
plt.xlabel("Flight Number", fontsize=20)  
plt.ylabel("Launch Site", fontsize=20)  
plt.show()
```



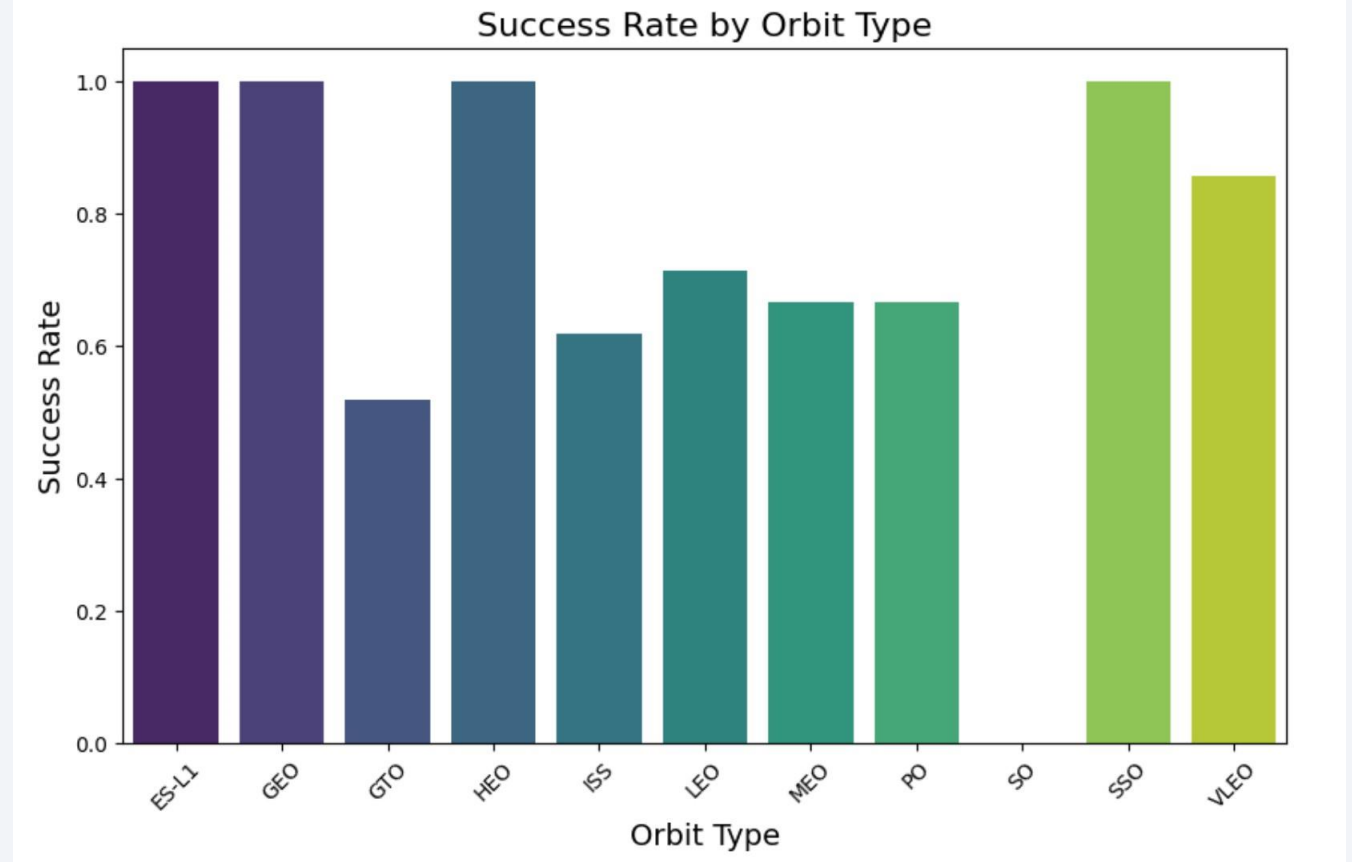
Payload vs. Launch Site

- Higher payload ($> 12000\text{kg}$) at CCAFS SLC 40 seem to have higher success rate.
- Lower payload ($< 5000\text{kg}$) at KSC LC 39A seem to have higher success rate.
- There are few data at VAFB SLC 4E.



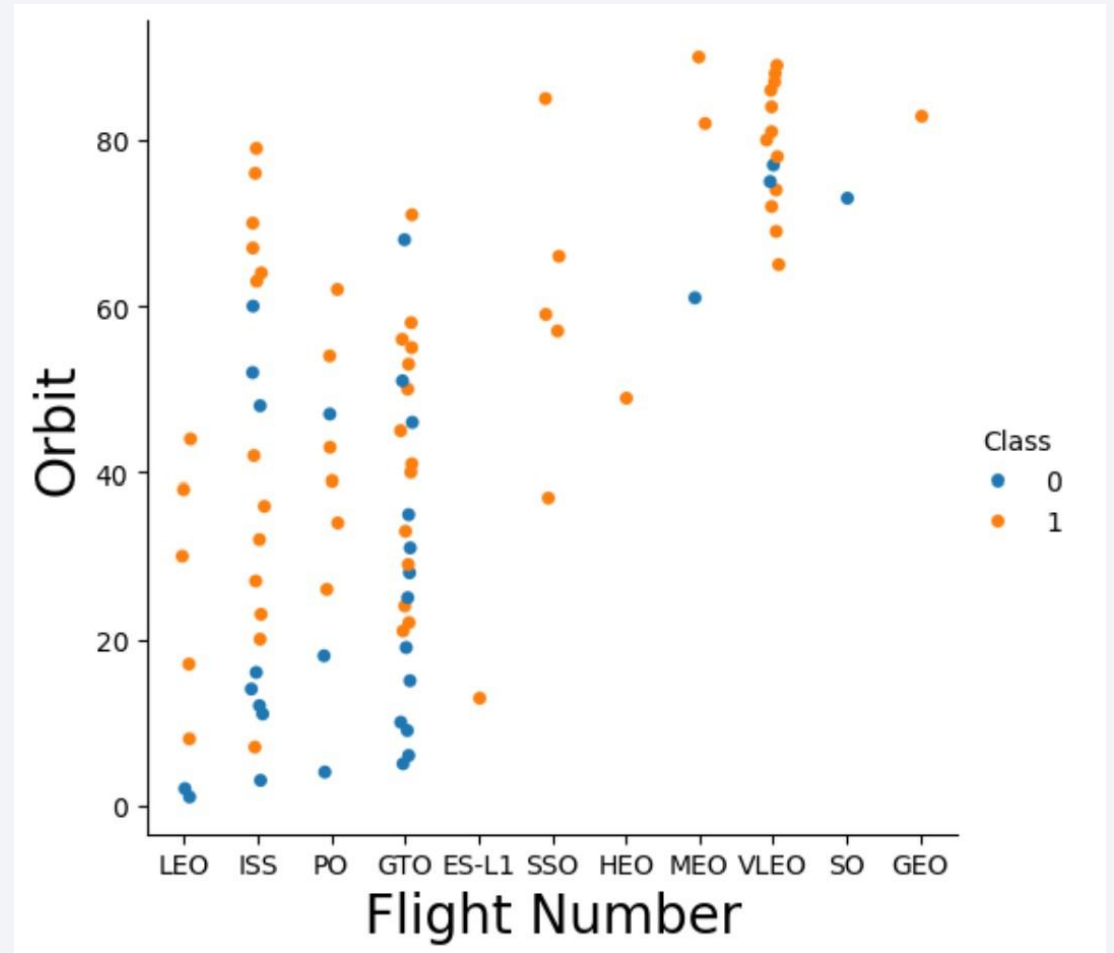
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have 100% success rates.
- SO has 0% success rate.
- Other orbit types have 50-80% success rates.



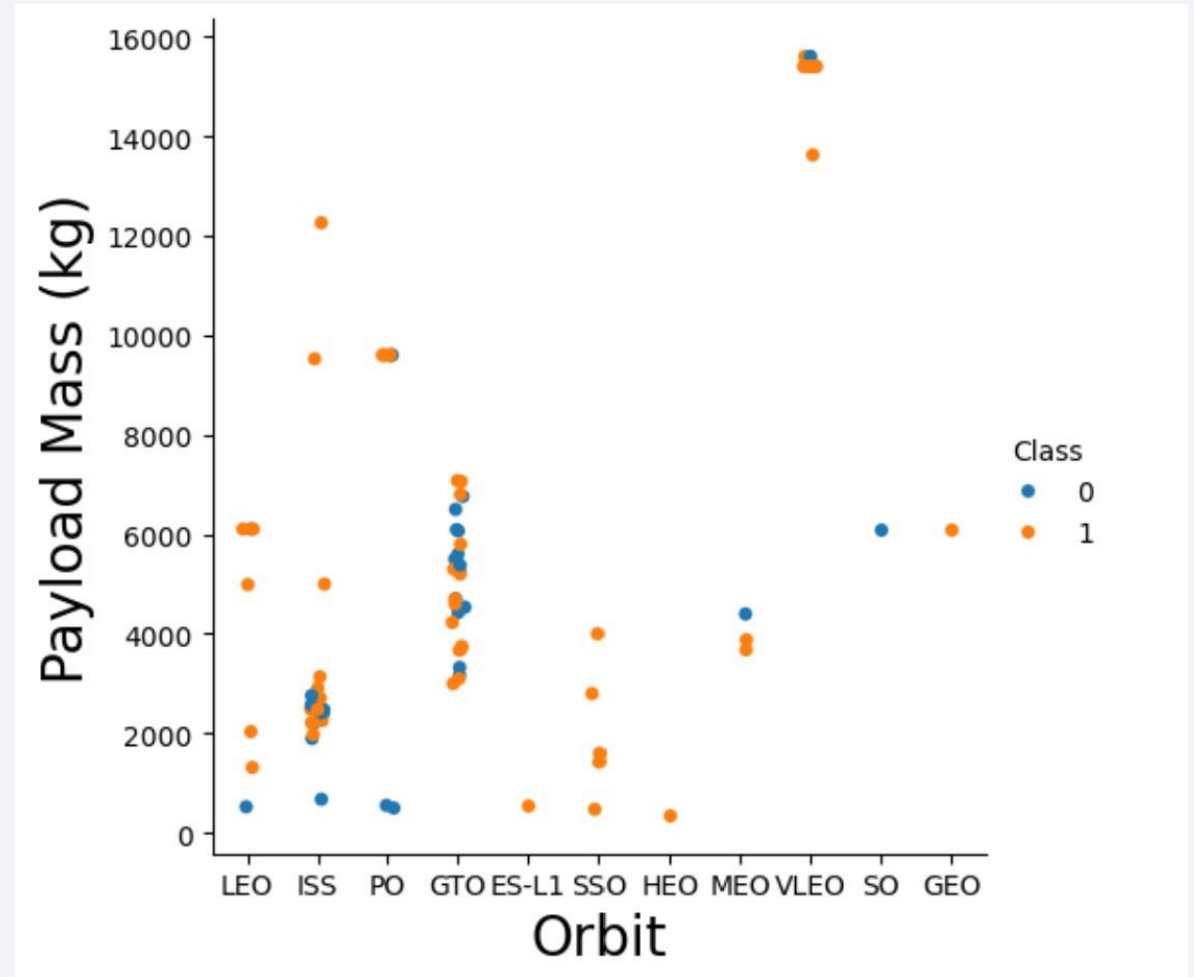
Flight Number vs. Orbit Type

- LEO, ISS, PO and GTO are the first orbit types being tested.
- Most tests were done on ISS and GTO compared to other orbit types.
- VLEO was more frequently tested later
- Other orbit types are infrequently tested.
- GTO seemed to have more failure tests compared to other orbit types.



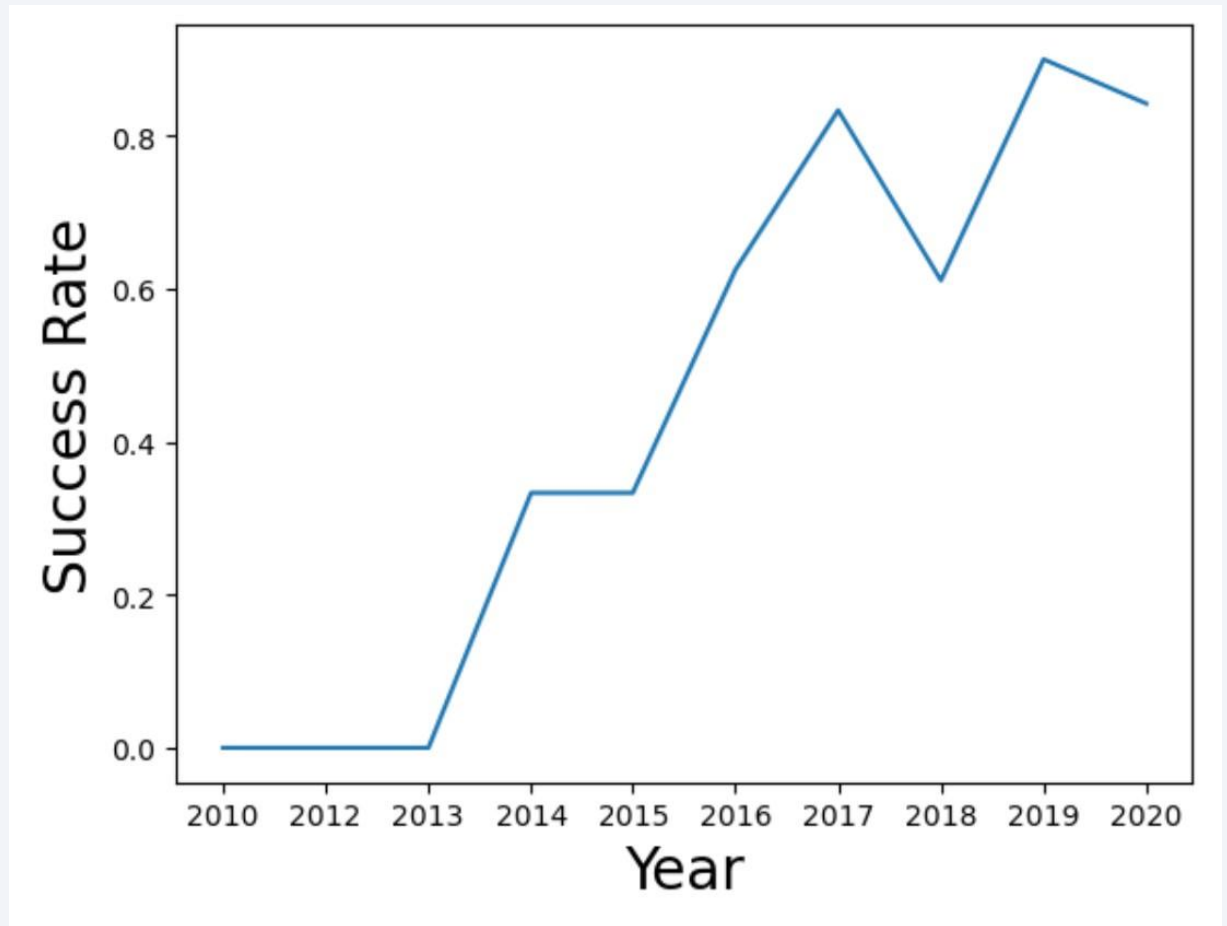
Payload vs. Orbit Type

- In LEO, ISS and PO, lower payload mass seemed to be associated with failure.
- In ISS and GTO, there seemed to be no distinct difference in payload mass between success and failure events.
- Highest payload tests were conducted in VLEO.



Launch Success Yearly Trend

- The success rate start to improve from 2013 to 2017.
- There seemed to be some setbacks in 2018 (to be determined why).
- Overall, there is an increasing trend in success rate between 2013 and 2020.



All Launch Site Names

- Unique launch sites:

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

```
import sqlite3

# Connect to the SQLite database
conn = sqlite3.connect('my_data1.db')

# SQL query to fetch unique launch sites
query = "SELECT DISTINCT Launch_Site FROM SPACEXTBL;"

# Execute the query and fetch the results
cursor = conn.cursor()
cursor.execute(query)
unique_launch_sites = cursor.fetchall()

# Close the connection
conn.close()

# Display the results
for site in unique_launch_sites:
    print(site[0])
```

Launch Site Names 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 *\nFROM SPACEXTBL\nWHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

Total Payload Mass

45,596 kg

%%sql

```
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
FROM SPACEXTBL
WHERE Customer LIKE 'NASA (CRS)';
```


Average Payload Mass by F9 v1.1

2,928.4 kg

%%sql

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

First Successful Ground Landing Date

2015-12-22

```
%%sql
```

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

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

%%sql

```
SELECT Booster_Version, payload
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
      AND PAYLOAD_MASS_KG_ > 4000
      AND PAYLOAD_MASS_KG_ < 6000;
```

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Total_Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

%%sql

```
SELECT Mission_Outcome, COUNT(*) AS Total_Count
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

Boosters Carried Maximum Payload

- 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

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

2015 Launch Records


Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

```
%%sql
SELECT
    SUBSTR(Date, 6, 2) AS Month,
    DATE,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)'
    AND SUBSTR(Date, 1, 4) = '2015';
```

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

Landing_Outcome	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

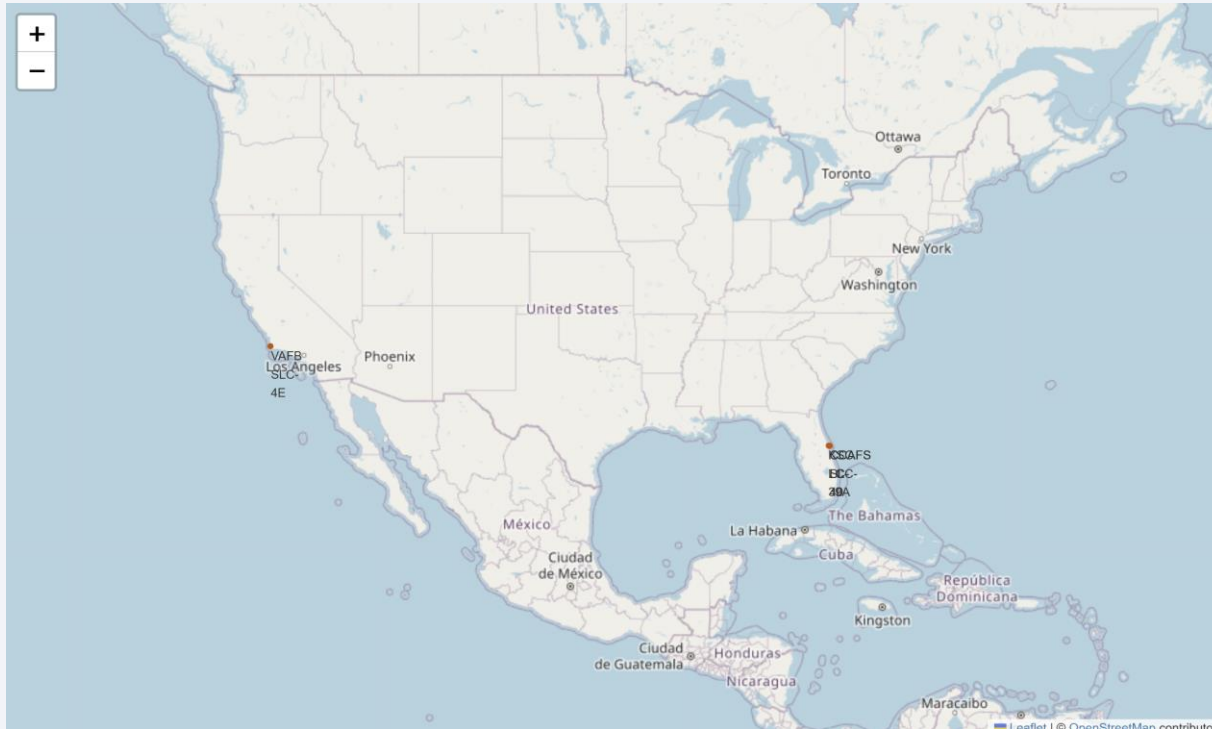
```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Outcome_Count
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

Section 3 Launch Site Proximity Analysis

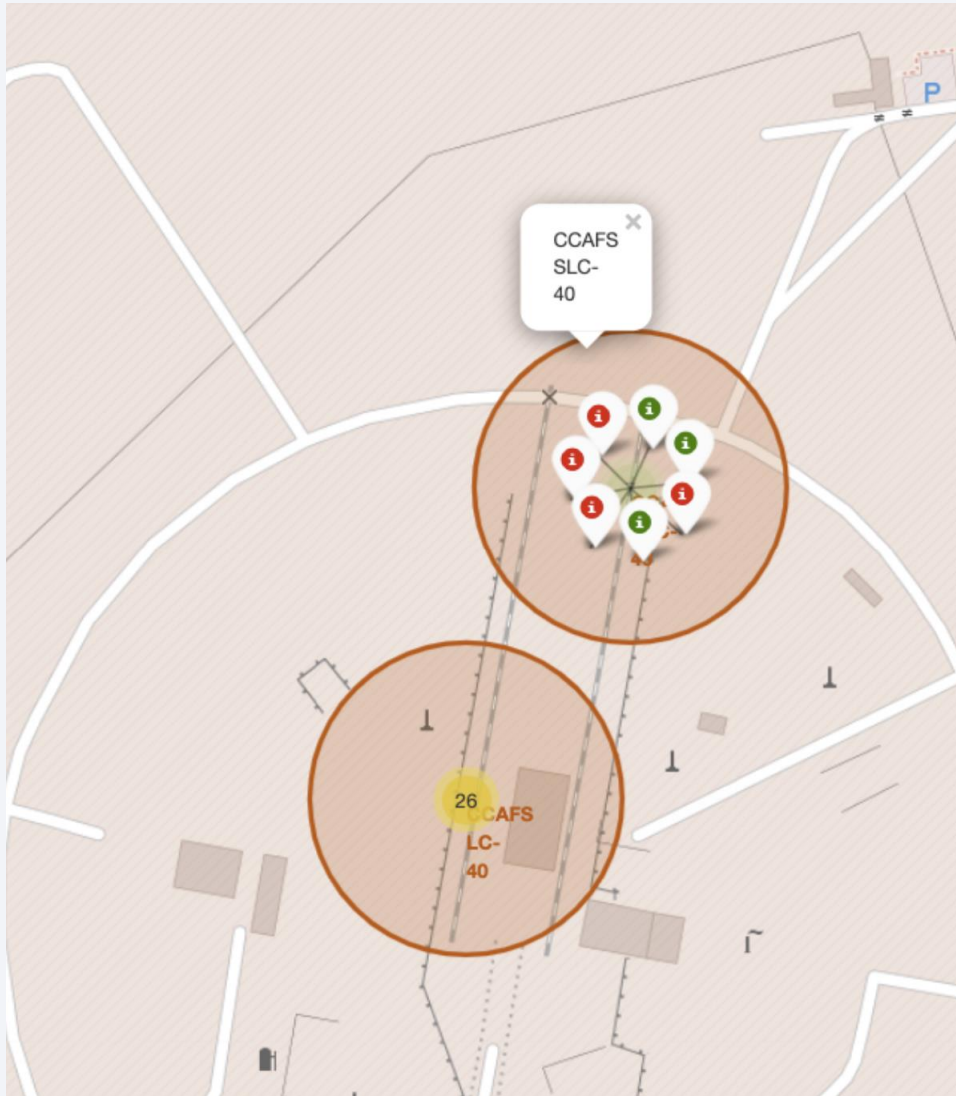


Launch Sites with Markers



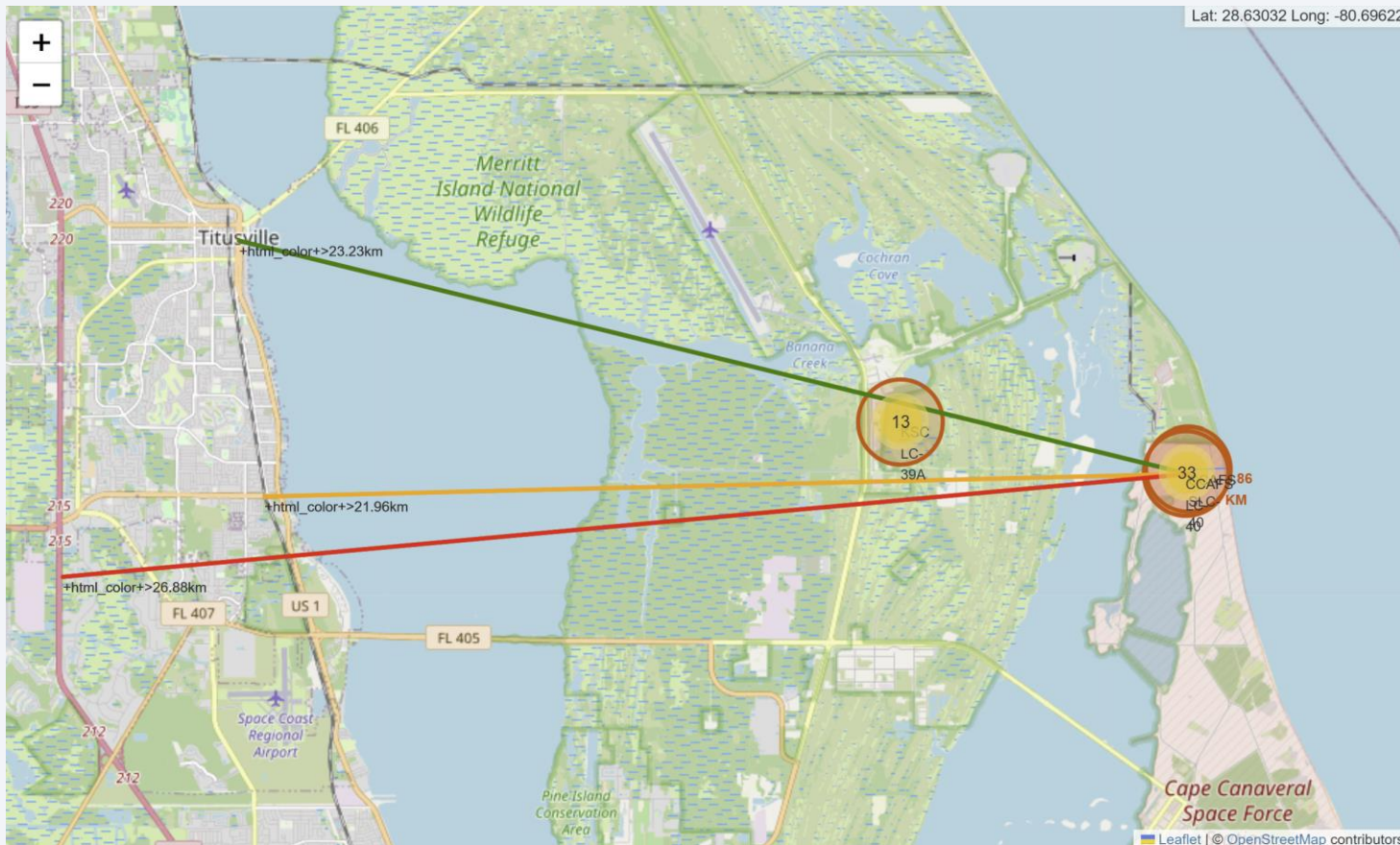
Proximity to the Equator: Launch sites closer to the equator have significant advantages for reaching equatorial orbits. Earth's rotation provides a natural boost for prograde orbits, making launches more efficient. Rockets launched near the equator benefit from the planet's higher rotational speed at these latitudes, reducing the need for additional fuel and boosters, ultimately lowering costs.

Launch Outcomes



- Green: successful launches
- Red: unsuccessful launches
- For example, launch site CCAFS SLC-40 has 3/7 successful launches.

Distance to Proximity



From CCAFS SLC-40 launch center:

Coastline (0.86km): Launch centers are often located near coastlines to reduce the risk to populated areas. In case of launch failures, debris or rockets falling back to Earth are more likely to land in the ocean, minimizing harm to people and infrastructure.

City (23.23km): Maintaining a safe distance from cities ensures that potential hazards, such as explosions, shockwaves, or falling debris, do not threaten urban populations.

Railway (21.96km) and Highway (26.88km): Launch centers rely on railways and highways to transport heavy and oversized rocket components, fuel, and equipment. Proximity to these transportation networks ensures efficient delivery and reduces logistical costs.

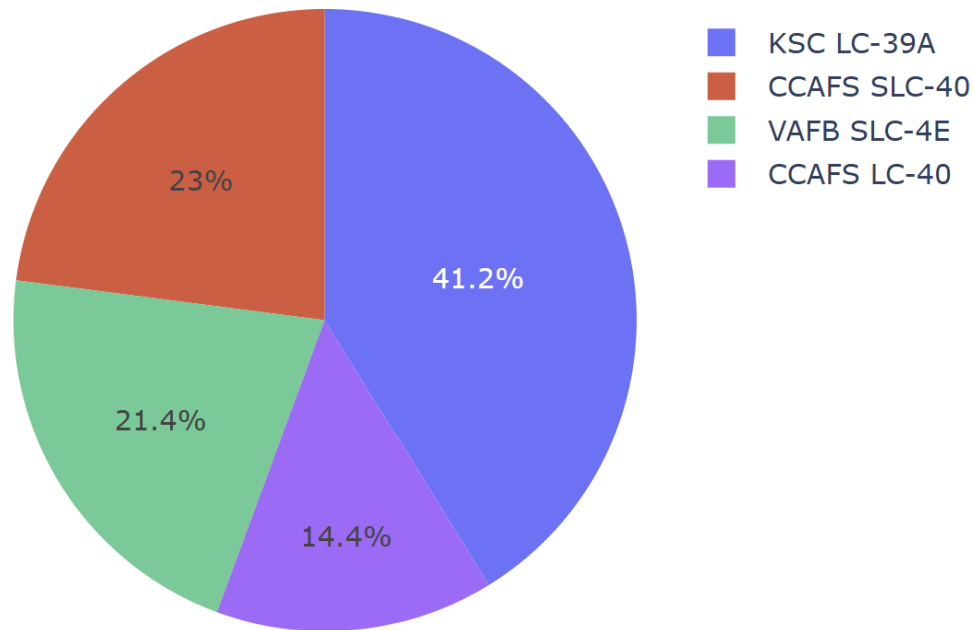
An anime-style illustration of a young person with short, light blue hair and bright blue eyes. They are wearing a white and grey futuristic flight suit with blue glowing accents. They are seated in a spaceship cockpit, with a large window behind them showing a view of Earth from space. The cockpit interior includes a steering wheel and various control panels with glowing blue lights.

Section 4

Building a Dashboard with Plotly Dash

Proportion of Successful Launches by Site

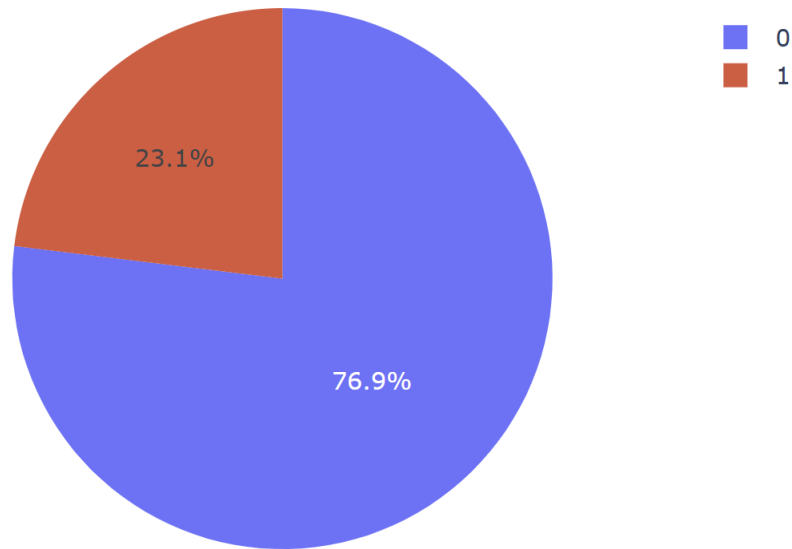
Proportion of Successful Launches by Site



- KSC LC-39A launch site has the highest proportion of successful launches
- CCAFS LC-40 launch site has the lower proportion of successful launches

Proportion of Successful Launches at KSC LC-39A

Successful and Unsuccessful Launches for Site KSC LC-39A




- At KSC LC-39A launch site, the successful rate of launches is 76.9%

Correlation between Payload Mass and Successfulness for All Sites



- Most successful launches have a payload between 2000 to 5000 kg
- However, this also could be due to the fact that most launches have payload between this range
- Further analysis is required



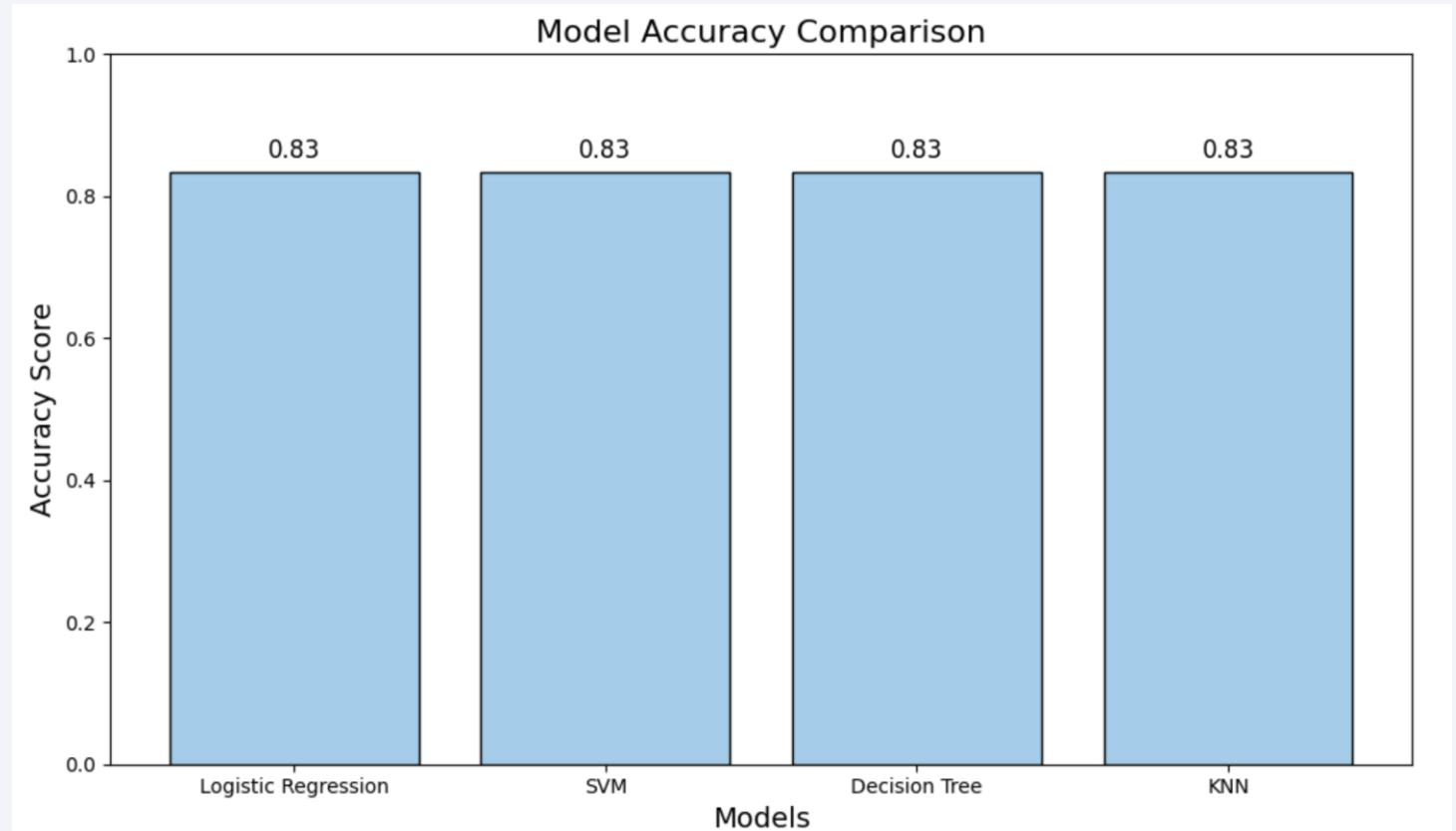
Section 5 Predictive Analysis (Classification)



Classification Accuracy

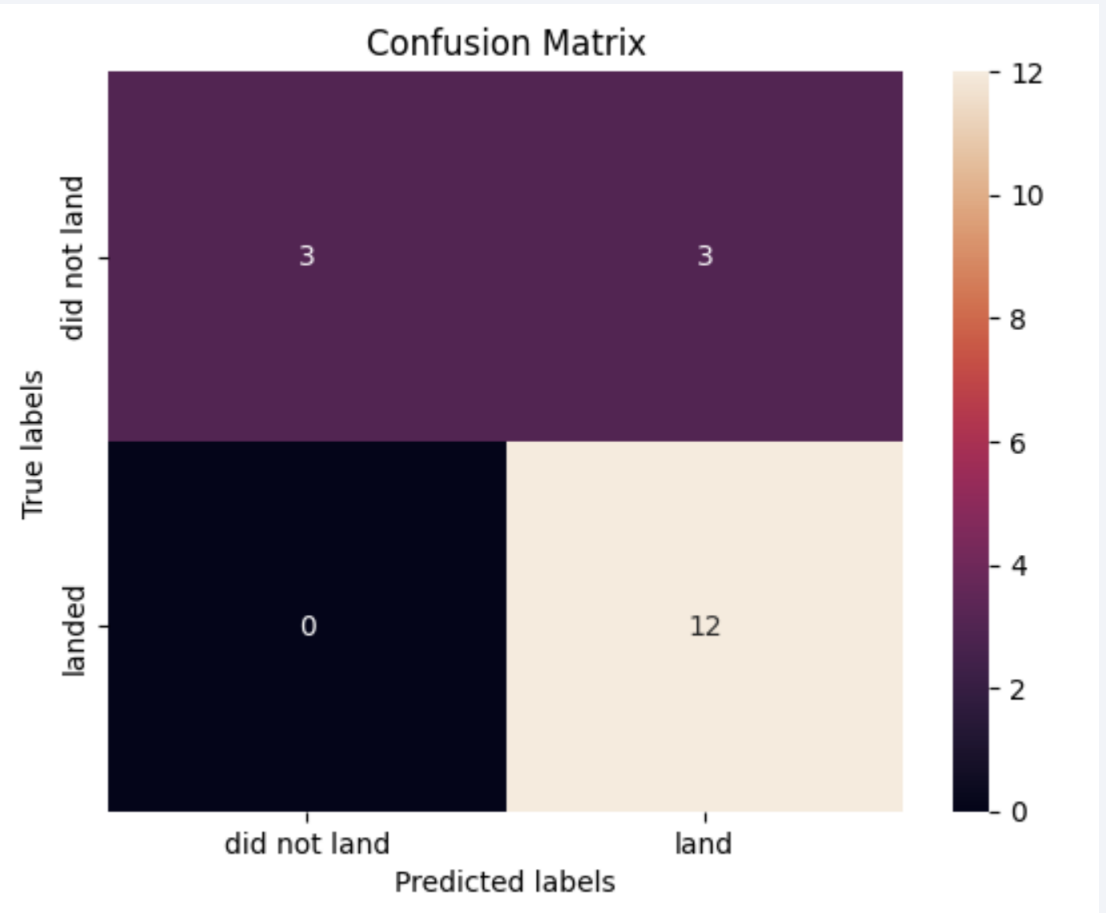
- All models have the same values for Jaccard Score, F1 Score and accuracy.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333



Confusion Matrix

- All the confusion matrix are the same, as all the model performed the same.
- There were:
 - 12 True positives
 - 3 True negatives
 - 3 False Positives
 - 0 False Negatives



Conclusions

- Later launches tend to be successful compared to the earlier launches, suggesting that SpaceX has gained experience to ensure higher success rate.
- Higher successful rate seems to be associated with particular payload at specific launch sites. However, the results may be bias due to small sample size at particular payload range.
- Launch sites are located near the equator to take advantage of the earth's rotation to provide natural boost.
- Success rate seems to be associated with orbit type, payload, launch site.
- Orbit type ES-L1, GEO, HEO, SSO have 100% success rate.
- KSC LC-39A has the highest success rate.
- More data is needed for the predictive analytics.

Thank you!

