IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Robinson Quintero Mesa
04/01/2025

Outline

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

Executive Summary

Summary of methodologies

•Data collection using the SpaceX REST API and web scraping techniques.

•Data processing to create a variable indicating success or failure.

•Data exploration through visualization, considering factors such as payload, launch site, flight number, and annual trends.

•Data analysis using SQL to calculate statistics such as total payload, payload range for successful launches, and total number of successful and failed launches.

•Exploration of success rates by launch site and their proximity to key geographical markers.

•Visualization of the most successful launch sites and the most successful payload ranges.

•Development of predictive models for landing outcomes using logistic regression, support vector machine (SVM), decision tree, and k-nearest neighbors (KNN).

Executive Summary

Summary of all results

Exploratory Data Analysis:
•Launch success has improved over time.
•KSC LC-39A has the highest success rate among launch sites.
•The ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.

Visualization/Analysis:
•Most launch sites are located near the equator, and all are close to the coast.

Predictive Analysis:
•All models performed similarly on the test set, although the decision tree model showed slightly better performance.

Introduction

Project background and context

SpaceX, a prominent player in the aerospace sector, is dedicated to making space travel more accessible and affordable. The company has made significant advancements, such as launching spacecraft to the International Space Station, deploying a satellite network to provide global internet connectivity, and conducting crewed missions to space. A key factor in SpaceX's cost-effectiveness is its ability to reuse the first stage of its Falcon 9 rocket, which brings the cost of each launch down to approximately 62 million dollars. In contrast, other companies without the capability to reuse the first stage face costs exceeding 165 million dollars per launch. By analyzing whether the first stage successfully lands, we can gauge the cost of a launch. This can be achieved by leveraging publicly available data and machine learning techniques to predict the likelihood of reusing the first stage.

Introduction

Problems you want to find answers

•The impact of payload weight, launch location, number of previous flights, and orbital destinations on the likelihood of a successful landing.

•Trends in the success rate of landings over time.

•Identifying the most effective machine learning model for predicting a successful landing (binary classification).

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

Steps

1.Request SpaceX Launch Data from API:
- •Send a request to the SpaceX API to retrieve rocket launch data.

2.Decode the Response:
- •Use .json() to decode the response and convert it into a data format that can be further processed, then transform it into a DataFrame using .json_normalize().

3.Request Launch Information via Custom Functions:
- •Implement custom functions to gather data from the SpaceX API.

4.Create a Dictionary from the Data:
- •Store the retrieved data in a dictionary for easier handling and manipulation.

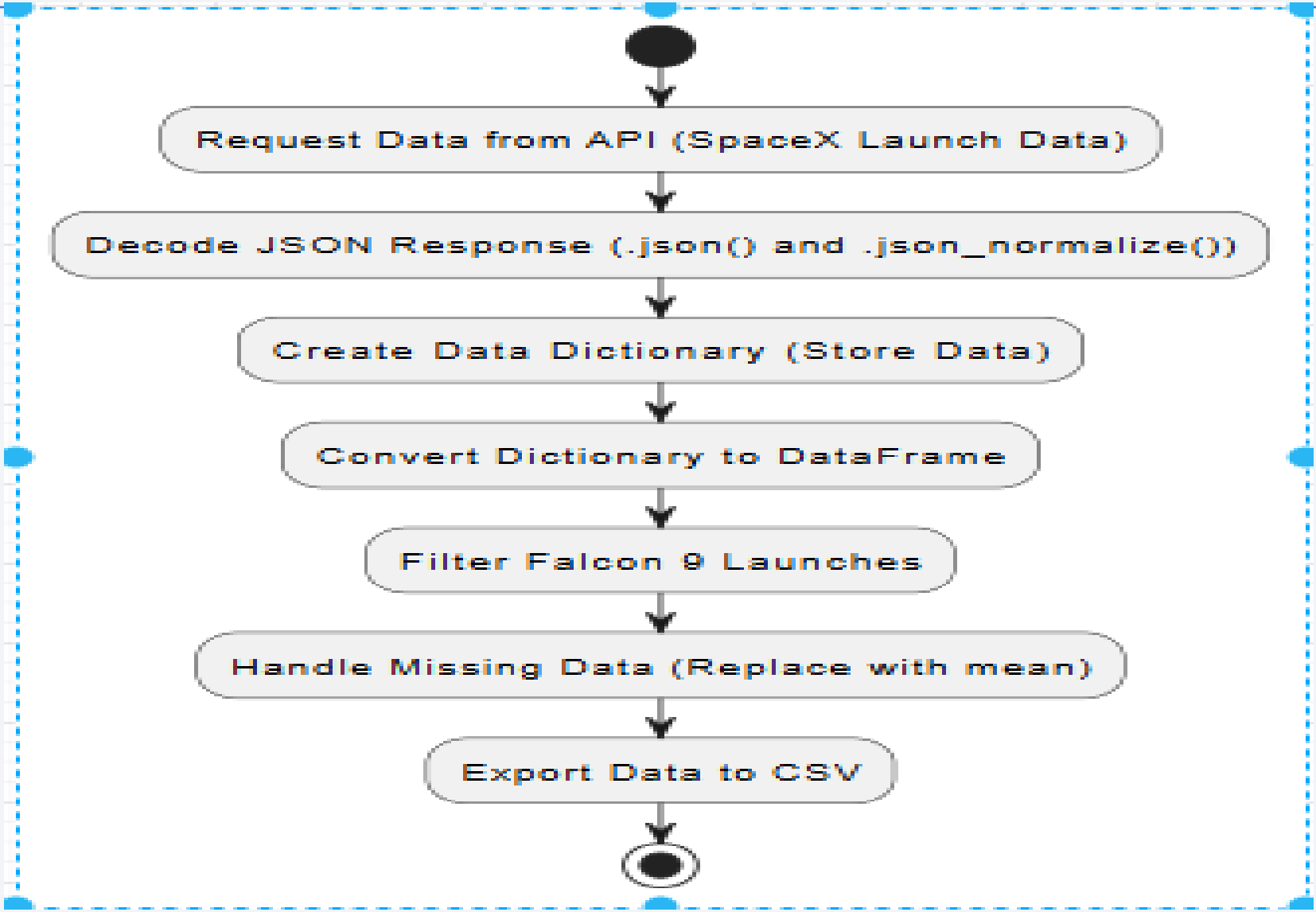
5.Convert the Dictionary to a DataFrame:
- •Convert the dictionary into a structured DataFrame for further analysis.

6.Filter the DataFrame for Falcon 9 Launches:
- •Filter the DataFrame to only include launches related to the Falcon 9 rocket.

7.Handle Missing Payload Mass Values:
- •Replace any missing values for payload mass with the mean value of the existing data, calculated using .mean().

8.Export Data to CSV:
- •Export the final cleaned and filtered data into a CSV file for further use.

Data Wrangling



EDA with Data Visualization

Visualizations

1.Flight Number vs. Payload Mass:

- Scatter plot to identify trends between flight number and payload mass.

2.Flight Number vs. Launch Site:

- Bar chart to compare the number of launches from different sites.

3.Payload Mass (kg) vs. Launch Site:

- Box plot to show the distribution of payload mass across launch sites.

4.Payload Mass (kg) vs. Orbit Type:

- Box plot or bar chart to compare payload mass across different orbit types.

Analysis

•Scatter Plots:

- Useful to visualize relationships between continuous variables. If a relationship exists, these variables may be important for predictive models.

•Bar Charts:

- Ideal for comparing discrete categories like launch sites and orbit types, highlighting key differences.

EDA with SQL

Queries

1.Show:

- **Unique launch site names**

  Query to get a list of distinct launch sites.

- **5 records where the lau nch site starts with 'CCA'**

  Show the first 5 records where the launch site name starts with 'CCA'.

- **Total payload mass carried by rockets launched by NASA (CRS)**

  Query to calculate the total payload mass carried by rockets launched by NASA.

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

  Query to get the average payload mass carried by the F9 v1.1 booster version.

EDA with SQL

2.List:

- **Date of the first successful landing on land platform**

  Query to get the date of the first successful landing on a land platform.

- **Names of boosters that successfully landed on an unmanned ship with a payload mass greater than 4000 but less than 6000**

  List of boosters that successfully landed on an unmanned ship with payload mass within the specified range.

- **Total number of successful and failed missions**

  Query to count the total number of successful and failed missions.

- **Names of booster versions that have carried the maximum payload**

  List of booster versions that have carried the maximum payload.

- **Failed landing results on unmanned ships, their booster version, and launch site during the months of 2015**

  Query to show failed landing results on unmanned ships, along with the booster version and launch site during 2015.

- **Landing result count between 04-06-2010 and 20-03-2017 (descending order)**

  Query to count landing results between the specified dates, ordered in descending order.

Build an Interactive Map with Folium

Launch Site Markers

•**Blue Circle at NASA Johnson Space Center Coordinates**:

A blue circle was added at the NASA Johnson Space Center coordinates, with a pop-up label displaying its name using its latitude and longitude coordinates.

•**Red Circles at All Launch Site Coordinates**:

Red circles were added at all launch site coordinates, with a pop-up label displaying the launch site name using its latitude and longitude coordinates.

Launch Result Color Markers

•**Successful (Green) and Failed (Red) Launch Markers**:

Color-coded markers were added to indicate successful (green) and failed (red) launches at each site, highlighting which sites have high success rates.

Distances Between Launch Site and Proximities

•**Color-coded Lines Showing Proximity**:

Color-coded lines were added to show the distance from the CCAFS SLC-40 launch site to nearby coastlines, railroads, roads, and the closest city.

Build a Dashboard with Plotly Dash

•**Dropdown for Launch Sites**:
Allows users to select all or a specific launch site.

•**Pie Chart for Successful Launches**:
Displays the percentage of successful vs. failed launches.

•**Payload Mass Range Slider**:
Enables users to select the payload mass range.

•**Scatter Plot for Payload Mass vs. Success Rate by Booster Version**:
Shows the correlation between payload mass and launch success rate for each booster version.

Predictive Analysis (Classification)

- **Create NumPy Array from 'Class' Column**
- **Standardize Data with StandardScaler**
- **Split Data Using train_test_Split**
- **Optimize with GridSearchCV (cv=10)**
- **Apply GridSearchCV to Models**: Logistic Regression, SVC, Decision Tree, KNN
- **Calculate Accuracy with .score()**
- **Evaluate Confusion Matrix**
- **Determine Best Model: Jaccard, F1, Accuracy**

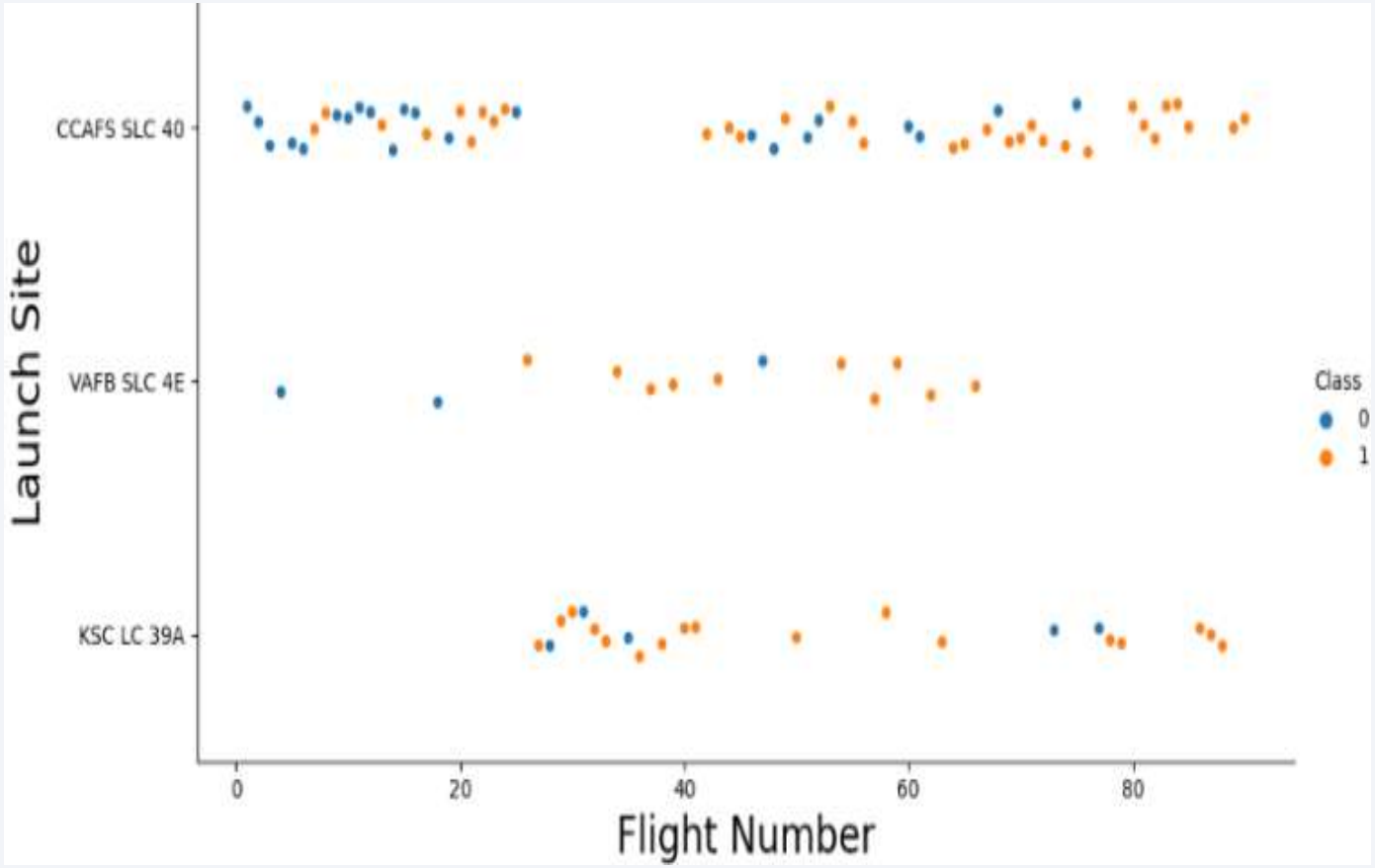
Results

- **Exploratory Data Analysis**
- Launch success has improved over time.
- KSC LC-39A has the highest success rate among landing sites.
- Orbits ES-L1, GEO, HEO, and SSO all have a 100% success rate.
- **Visual Analysis**
- Most launch sites are near the equator and close to the coast.
- Launch sites are sufficiently far from populated areas (cities, highways, railroads) to avoid damage from a failed launch, but close enough to transport personnel and materials.
- **Predictive Analysis**
- The decision tree model is the best predictive model for the dataset.


Section 2
Insights drawn from EDA

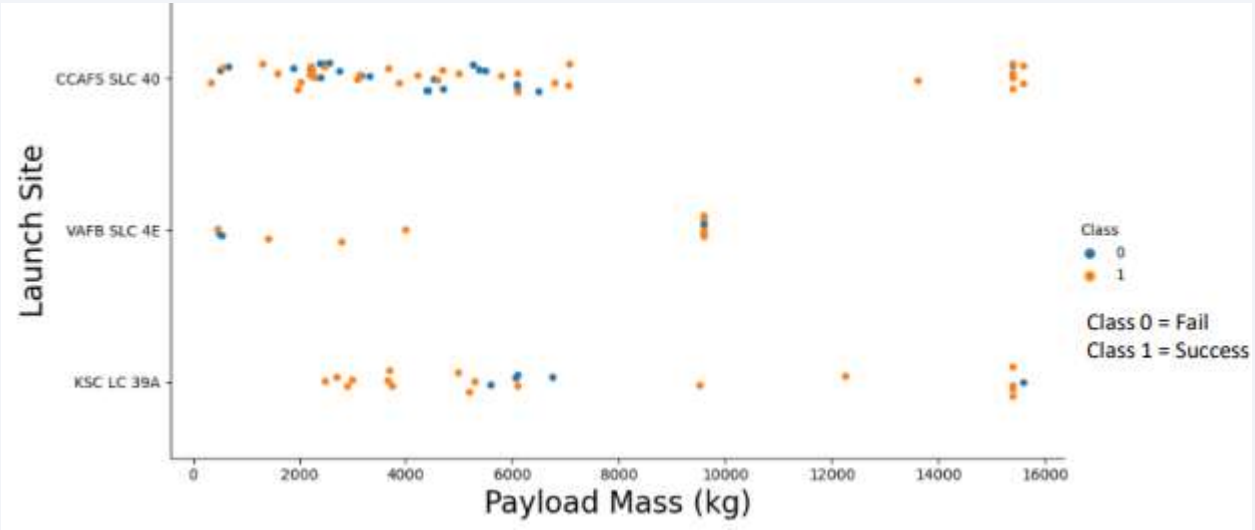
Flight Number vs. Launch Site

- **Exploratory Data Analysis**
- Earlier flights had a lower success rate (blue = failure).
- Later flights had a higher success rate (orange = success).
- About half of the launches were conducted from the CCAFS SLC 40 site.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be inferred that newer launches have a higher success rate.



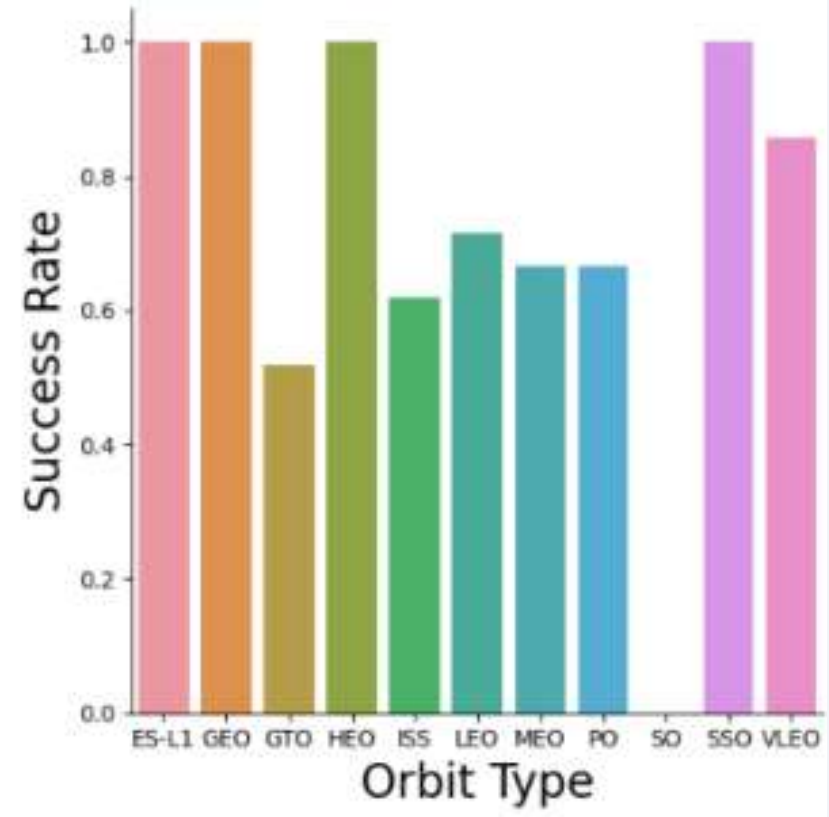
Payload vs. Launch Site

- **Exploratory Data Analysis**
- Generally, the higher the payload mass (kg), the higher the success rate.
- Most launches with a payload greater than 7,000 kg were successful.
- KSC LC 39A has a 100% success rate for launches with payloads under 5,500 kg.
- VAFB SKC 4E has not launched anything weighing more than approximately 10,000 kg.



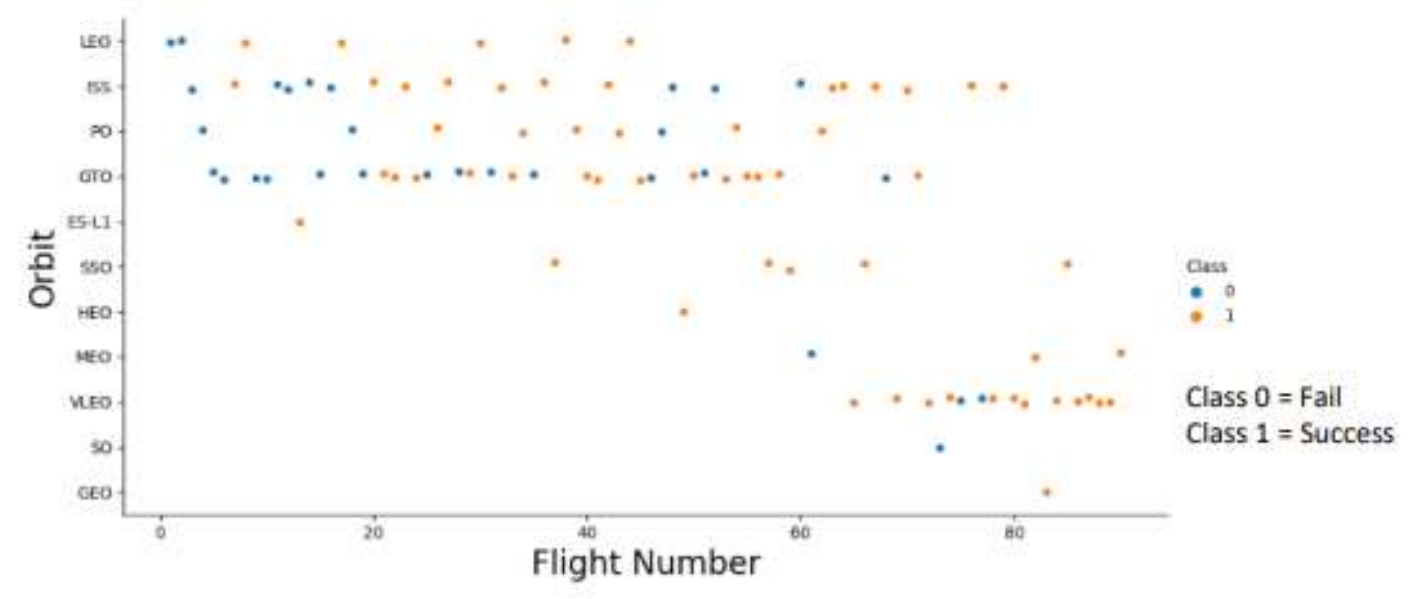
Success Rate vs. Orbit Type

- **Exploratory Data Analysis**
- **100% Success Rate**: ES-L1, GEO, HEO, and SSO.
- **50% to 80% Success Rate**: GTO, ISS, LEO, MEO, PO.
- **0% Success Rate**: SO.



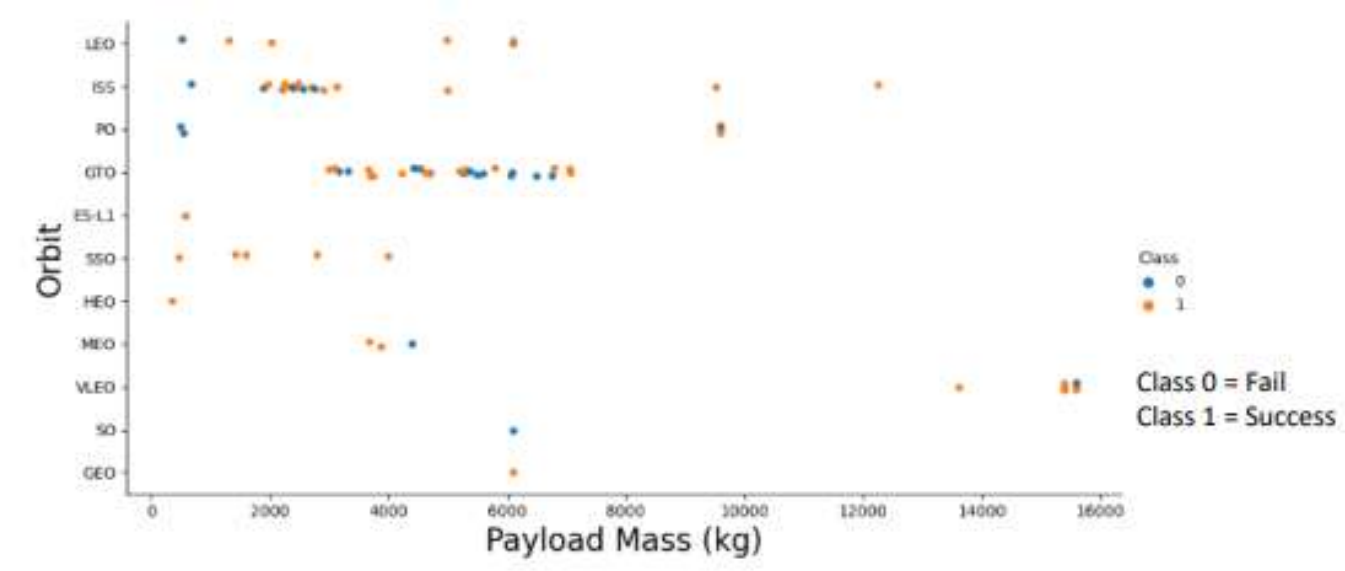
Flight Number vs. Orbit Type

- **Exploratory Data Analysis**
- Typically, the success rate improves as the number of flights increases for each orbit.
- This pattern is most noticeable in the LEO orbit.
- On the other hand, the GTO orbit does not follow this pattern.



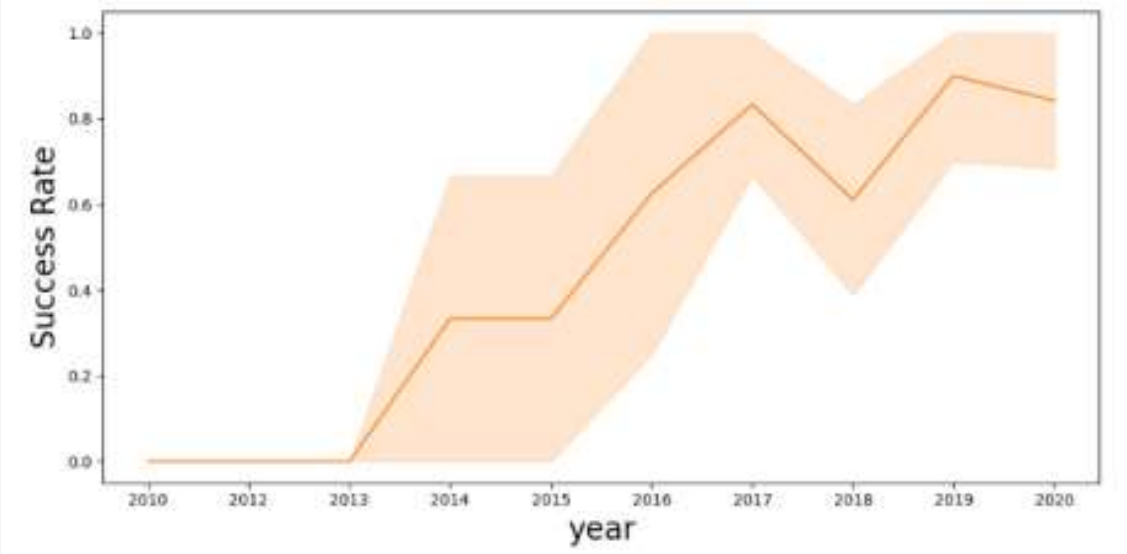
Payload vs. Orbit Type

- **Exploratory Data Analysis**
- Larger payloads tend to perform better in LEO, ISS, and PO orbits.
- The GTO orbit shows inconsistent success with heavier payloads.



Launch Success Yearly Trend

- **Exploratory Data Analysis**
- The success rate improved between 2013 and 2017, and again between 2018 and 2019.
- The success rate decreased between 2017-2018 and 2019-2020.
- Overall, the success rate has improved since 2013.



All Launch Site Names

launch sites

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

```
%sql SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9
   sqlite:///my_data1.db
Done.
```

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

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
   sqlite:///my_data1.db
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Launch Site Names Begin with 'CCA'

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

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

- **Total Payload Mass**
- 45,596 kg (total) carried by rockets launched by NASA (CRS).
- **Average Payload Mass**
- 2,928 kg (average) carried by the F9 v1.1 booster version.

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

1
45596

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

1
2928

Average Payload Mass by F9 v1.1

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

First Successful Ground Landing Date

- 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);
```

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

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

Successful Drone Ship Landing with Payload between 4000 and 6000

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Total Number of Successful and Failure Mission Outcomes

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

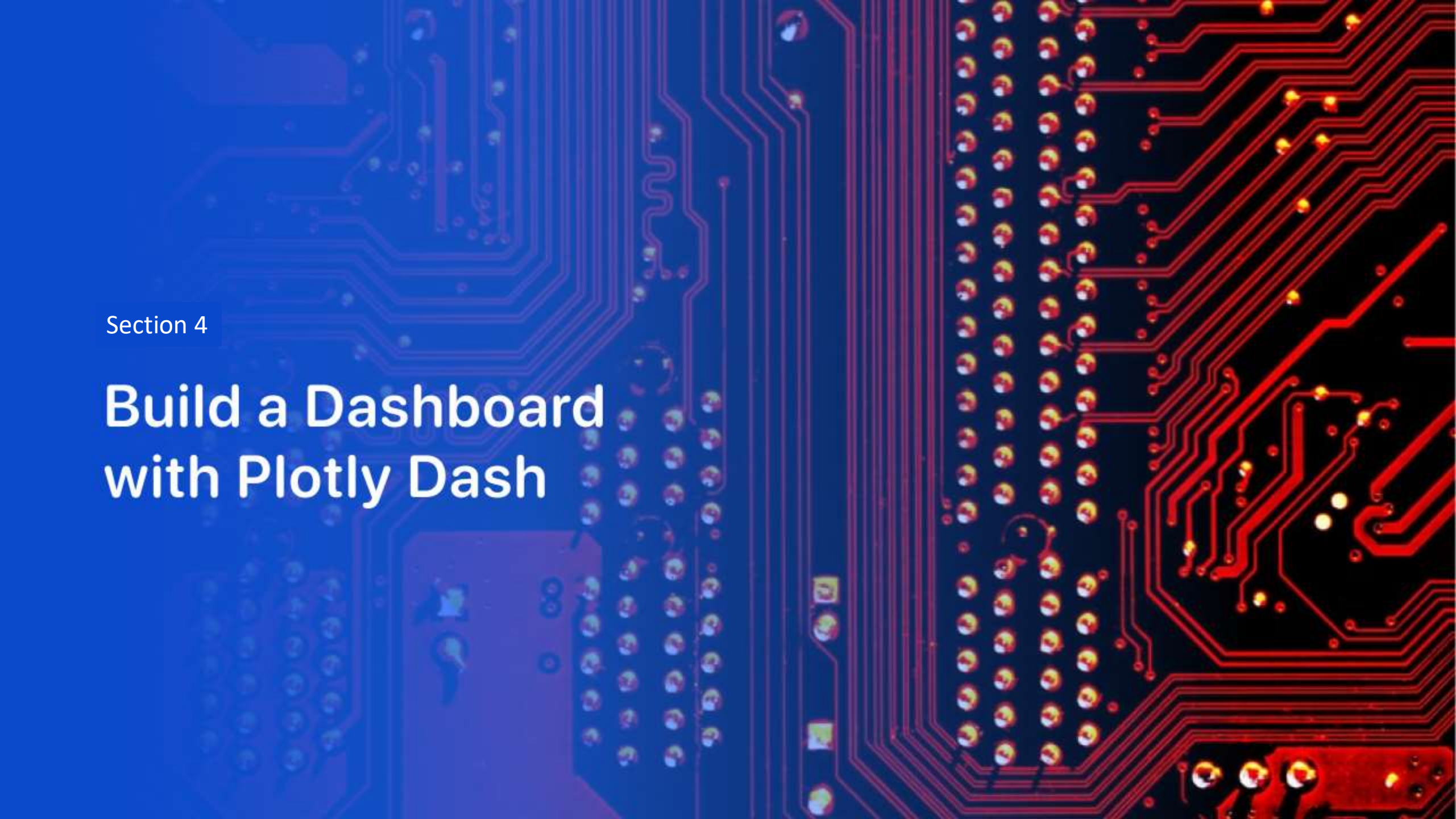
Section 3

Launch Sites Proximities Analysis

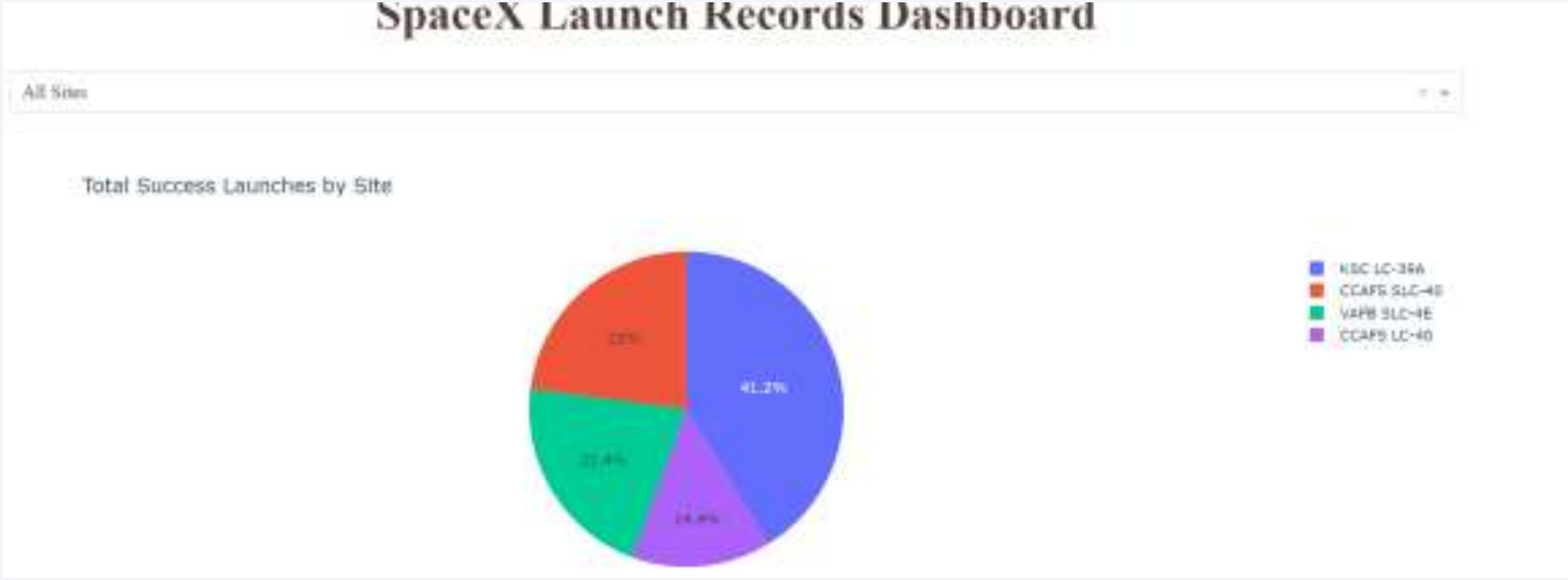
Launch Sites

Launch Outcomes

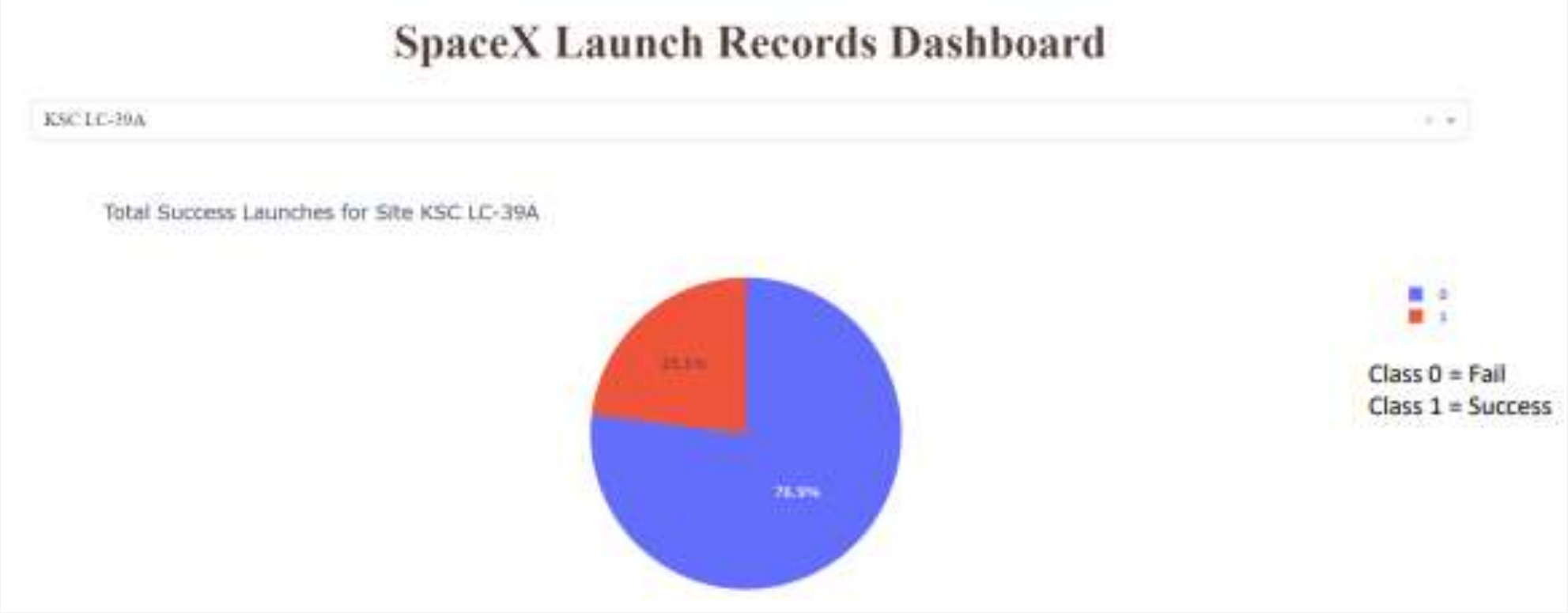
Distance to Proximities


Section 4
Build a Dashboard with Plotly Dash

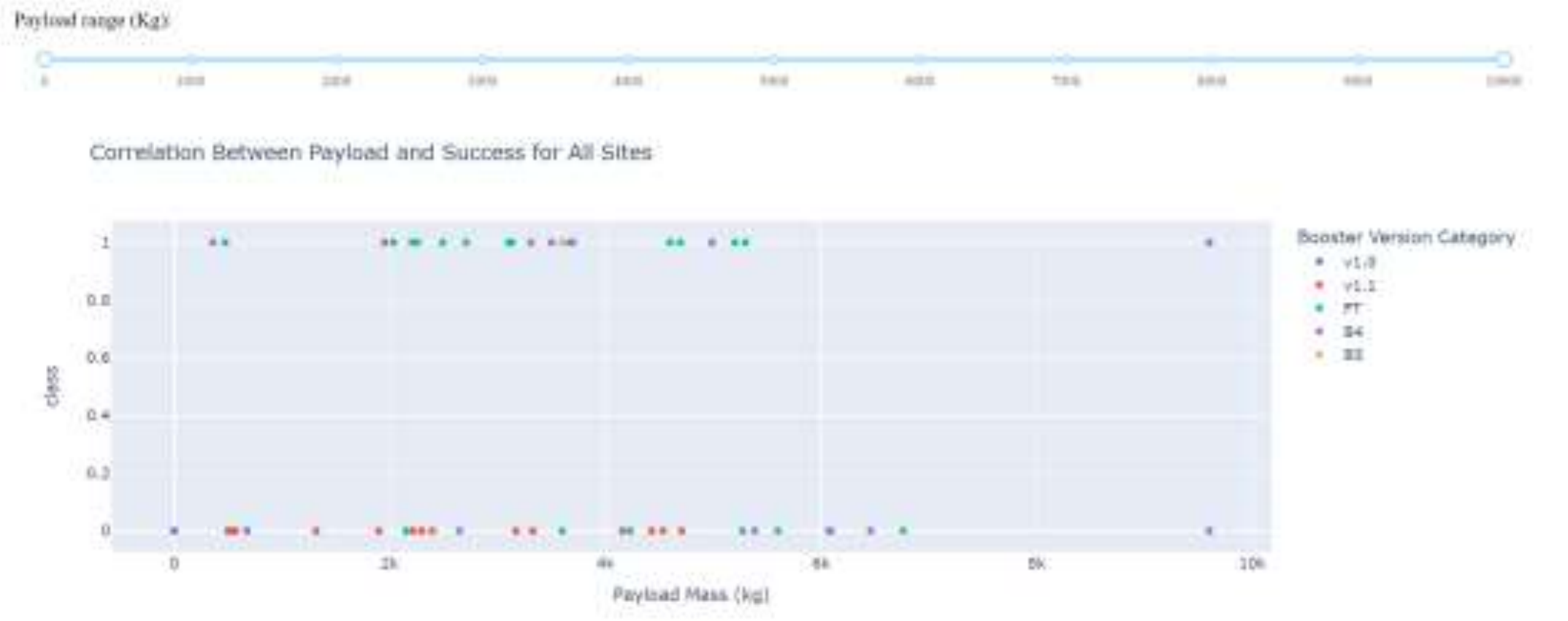
Launch Success by Site



Launch Success (KSC LC-29A)



Payload Mass and Success



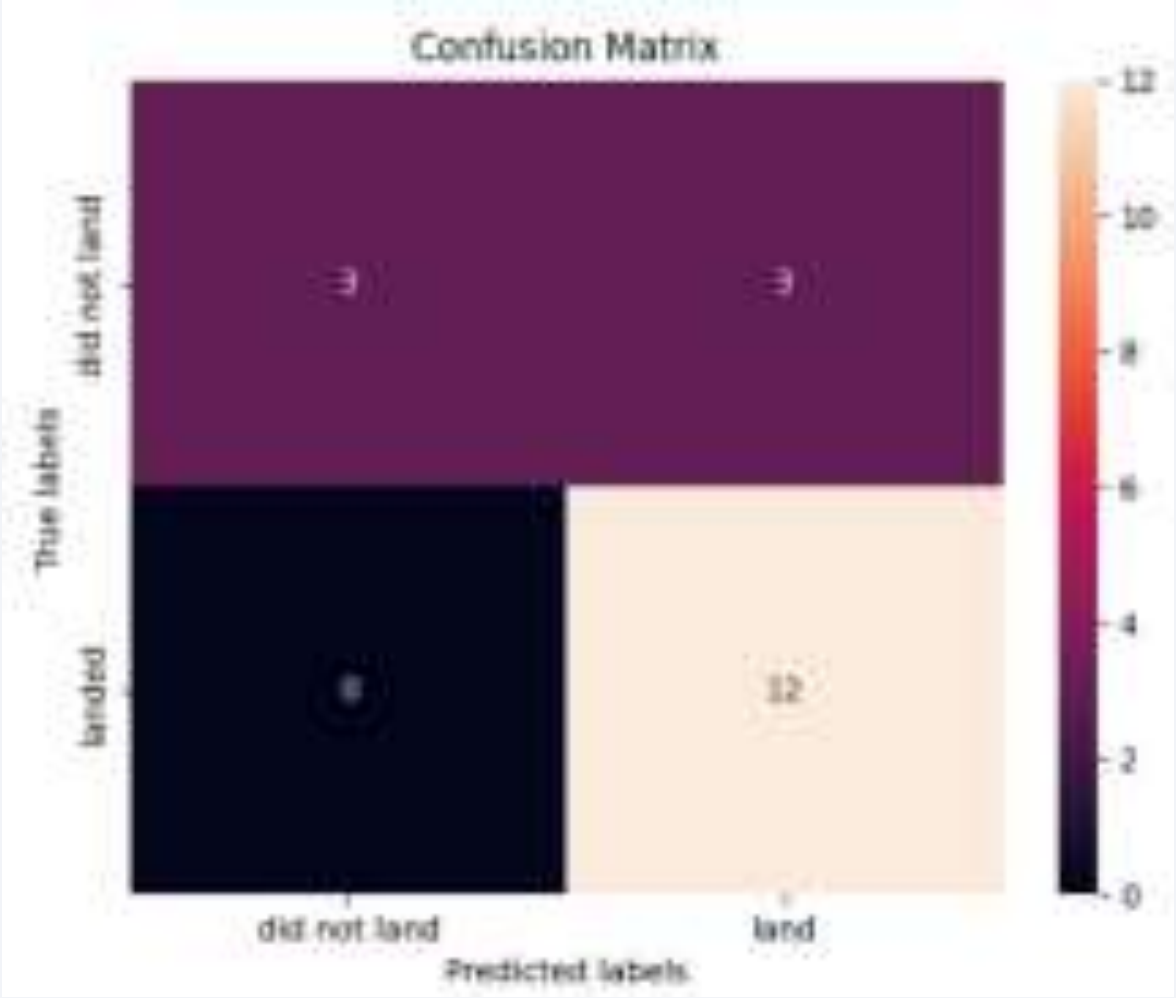
Section 5

Predictive Analysis (Classification)

Classification 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



Conclusions

- **Study Findings**
- **Model Accuracy**: The decision tree model slightly outperformed the others, though all models showed similar results on the test data.
- **Proximity to Equator**: Most launch sites are located near the equator, benefiting from Earth's rotational speed. This reduces the need for additional fuel and boosters, making launches more cost-effective.
- **Coastal Location**: All launch sites are situated near the coast.
- **Launch Success Rate**: Success rates have improved progressively over the years.
- **KSC LC-39A**: This site stands out with the highest success rate and a perfect track record for payloads under 5,500 kg.
- **Orbit Success Rates**: Orbits such as ES-L1, GEO, HEO, and SSO have a flawless success rate.
- **Payload Mass and Success**: In general, larger payloads tend to correlate with higher success rates across all launch sites.

Appendix

Through this analysis, I learned several key aspects that impact the success of space launches. By applying machine learning models, like decision trees, I understood how to compare and optimize predictive models. I discovered that the geographical location of launch sites, near the equator and coast, influences efficiency by reducing fuel costs.

Additionally, I observed that the success rate has improved over time, reflecting technological advancements and more effective strategies. I also noted that launches with heavier payloads tend to have higher success rates, suggesting more investment and preparation.

I used Python to process data, create models, and generate visualizations, and applied techniques like the confusion matrix to evaluate models. The charts and code snippets were also essential for interpreting results and conducting in-depth analysis.

Thank you!