Winning Space Race with Data Science

Michael Feranda
03/13/2024

Outline

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

Executive Summary

- In this project, we will predict if the SpaceX Falcon 9 first stage will land successfully.
- Methodologies:
  - Automated API Data Extraction
  - Modular Data Processing
  - Data Refinement & Structuring
  - Targeted Feature Engineering
- Summary of results
  - There is a correlation with the outcome of the launches.
  - Use of decision tree may be the best method for prediction.

Introduction

- Project background and context
  - SpaceX's reusable Falcon 9 rockets offer significant cost advantages in space launches. Understanding the factors that influence successful first-stage landings is crucial for competitive analysis and cost optimization.
- Problems you want to find answers
  - We aim to predict Falcon 9 first-stage landing success. This will enable us to determine launch costs more accurately, providing a strategic advantage for alternative launch providers.

Section 1

Methodology

Methodology

Executive Summary

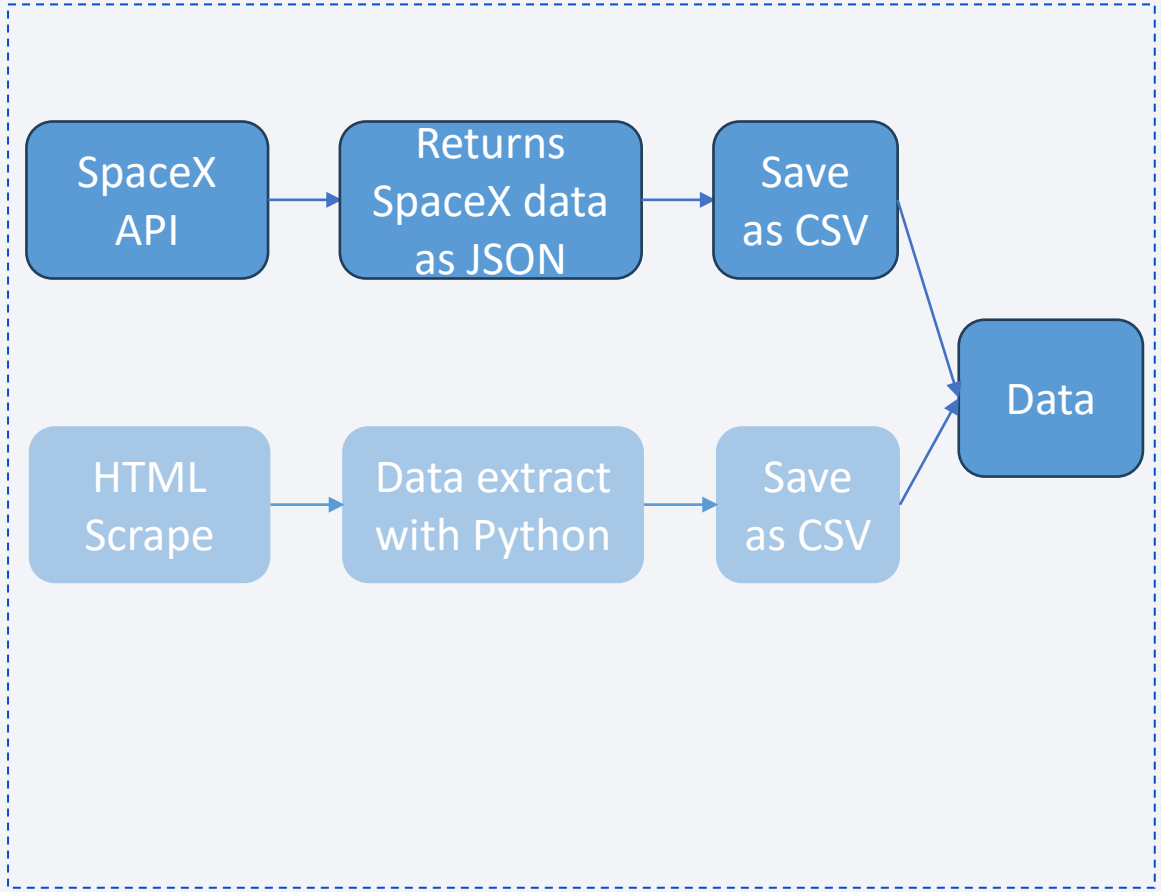
- Automated API Data Extraction:
  - Leveraged SpaceX's public API to gather comprehensive historical launch data.
  - Implemented structured functions to efficiently extract and organize data from various API endpoints.
- Data Refinement & Structuring:
  - Utilized Pandas to transform raw data into a clean, structured DataFrame for analysis. Implemented error handling to increase data integrity.
- Targeted Feature Engineering:
  - Focused on extracting and refining key variables relevant to Falcon 9 first-stage landing prediction with interactive visuals using Follium and Potly.
- Machine learning prediction:
  - Logistic regression, Decision tree, K-nearest neighbors (KNN), Support vector machine (SVM)

Data Collection

Our data collection began with accessing the SpaceX public API, the primary source for historical launch information. Using Python's requests library, we systematically retrieved data from various endpoints, including launches, rockets, launchpads, payloads, and cores. This allowed us to build a comprehensive dataset encompassing launch date, rocket details, launch site coordinates, payload mass, and core landing outcomes.

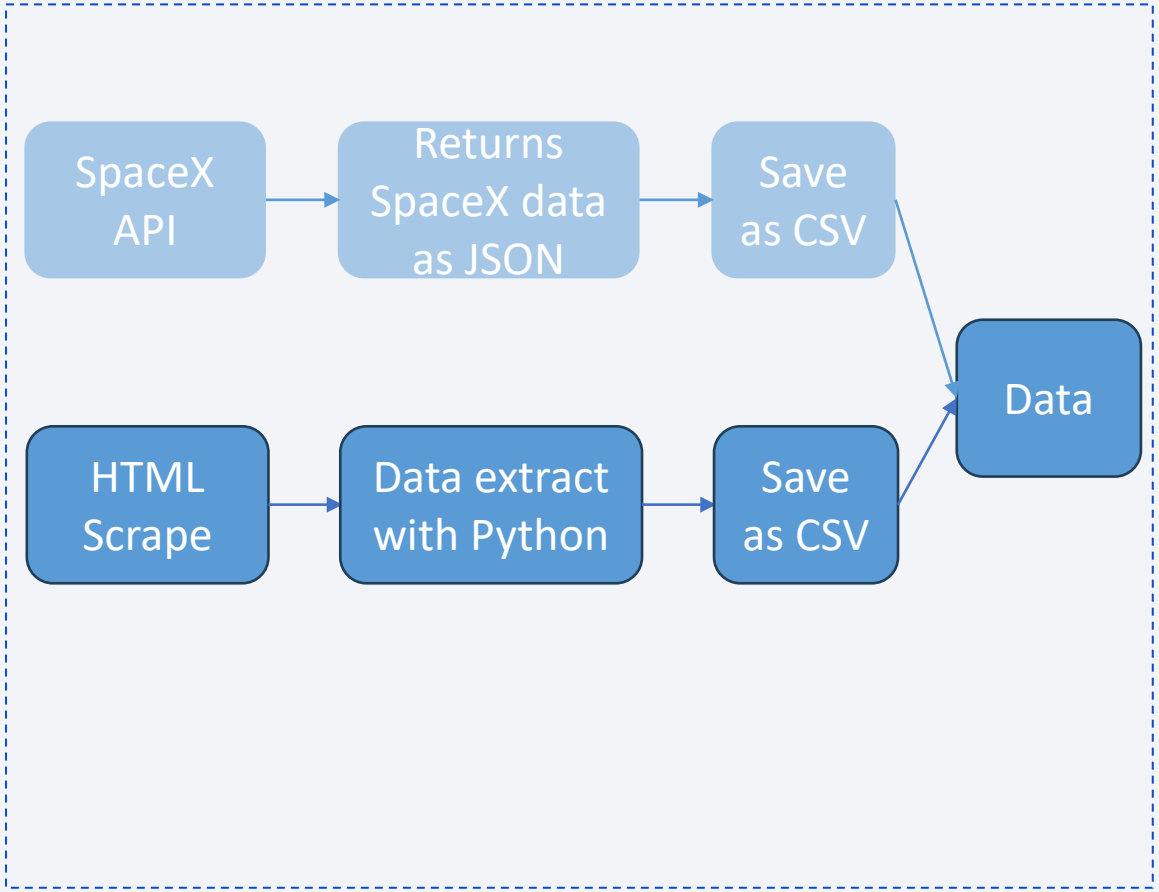
Data Collection – SpaceX API

- SpaceX data is gathered from the API as JSON.
  - This provides information about launches like rocket used, location data, payload, specifications, landing outcome.
- Data Collection Jupyter File



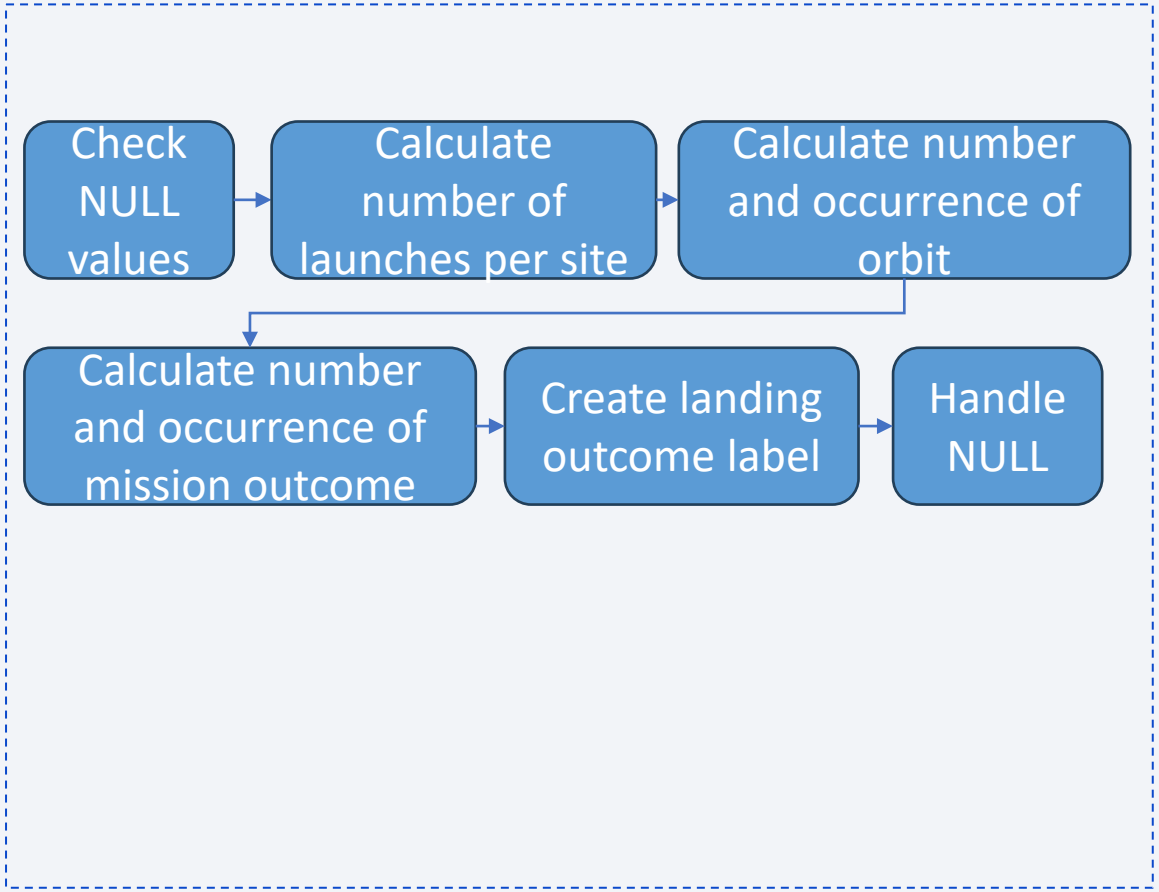
Data Collection - Scraping

- Data scraped from Wikipedia using BeautifulSoup
  - This provided information about launches like rocket used, location data, payload, specifications, landing outcome.
- Data Collection Jupyter File



Data Wrangling

- Loaded data using pandas, read_csv
- Identify missing values and calculate percentages.
  - Check data types (numerical, categorical).
  - Analyze launch sites (value_counts), orbits (value_counts), and mission outcomes (value_counts).
- Create Labels:
  - Define successful (1) and unsuccessful (0) landing outcomes based on mission outcomes.
  - Generate Class labels.Keywords: Outcome, Class, labeling, binary classification
- Data Collection Jupyter File



EDA with Data Visualization

- **Types of visualizations with EDA:**
  - Bar Chart: Visualizes the count of records for each launch site, providing insight into the distribution of data across different locations.
  - Pie Chart: Illustrates the proportion of successful and failed outcomes, giving a clear view of the overall success rate.
  - Scatter Plot: Depicts the relationship between payload mass and flight number, helping to identify any trends or correlations between these two variables.
  - Bar Chart (Outcome by Launch Site): Shows the success and failure counts for each launch site, allowing for a comparison of performance across different locations.
  - Histogram: Visualizes the distribution of payload mass, providing insights into the frequency of different payload mass ranges.
- Data Collection Jupyter File

EDA with SQL

- Launch Site Filtering: Queries focused on retrieving data related to specific launch sites or filtering based on launch site name patterns.
    - Payload Mass Analysis: Queries calculated statistics such as total, average, and maximum payload mass, often for specific subsets of the data (e.g., by booster type or organization).
    - Landing Outcome Analysis: Queries explored landing outcomes by:
        - Retrieving unique outcome values.
        - Ordering outcomes.
        - Identifying the date of the first successful ground pad landing.
        - Counting successful landings (overall or by type).
        - Filtering records based on success or failure and landing type.
- Data Collection Jupyter File

Build an Interactive Map with Folium

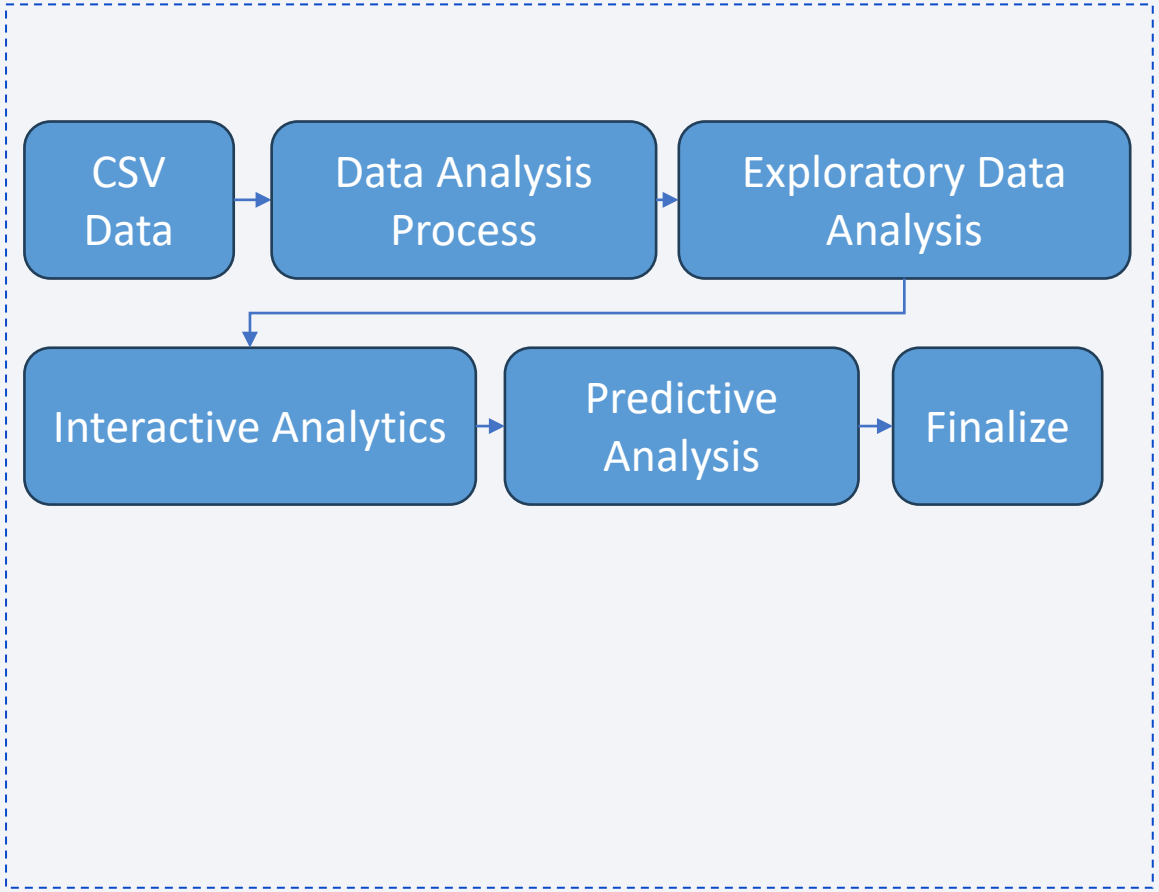
- Map Objects Created
  - Markers: Markers were added to the map to indicate the precise location of each launch site.
  - Circles: Circles were added to the map to highlight the area around each launch site,
- Explanation of Why Objects Were Added
  - Markers: These were used to pinpoint the exact geographical coordinates of each launch site, making it easy to visualize their distribution on the map.
  - Circles: These were added to provide a visual representation of the proximity or surrounding area of each launch site, helping to understand the geographical context of each site
- Data Collection Jupyter File

Build a Dashboard with Plotly Dash

- Plots/Graphs Added
  - Line Plot: A line plot is added to the dashboard.
- Interactions Added
  - Input Year: An input field allows the user to enter a year.
  - Callback: A callback function is defined to update the line plot based on the year entered by the user.
  - Explanation of Why Plots and Interactions Were Added
  - Line Plot: This plot is used to visualize the average arrival delay time by month for a given year, enabling users to observe trends and patterns in flight delays over the months.
  - Input Year: This interaction allows the user to dynamically change the year for which the average arrival delay time is displayed, providing the ability to explore how flight delays vary across different years.
  - Callback: The callback connects the input field with the line plot, ensuring that the plot updates automatically whenever the user enters a different year. This interactivity makes the dashboard more user-friendly and facilitates dynamic exploration of the data.
- Data Collection Jupyter File

Predictive Analysis (Classification)

- Data Loading and Preparation: Datasets were loaded using pandas, and preprocessing techniques were applied, including standardization. The data was split into training and testing sets.
- Model Selection and Training: Four classification algorithms were chosen: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors. Hyperparameter tuning was performed using GridSearchCV to find the best parameters for each model.
- Model Evaluation: The models were evaluated using the test data, and their accuracy was calculated using the score method. The model with the highest accuracy was determined to be the best-performing model.
- Model Improvement: The hyperparameters of each model were tuned to improve their performance. Different classification algorithms were compared to find the one that performs best on the test data.
- Data Collection Jupyter File



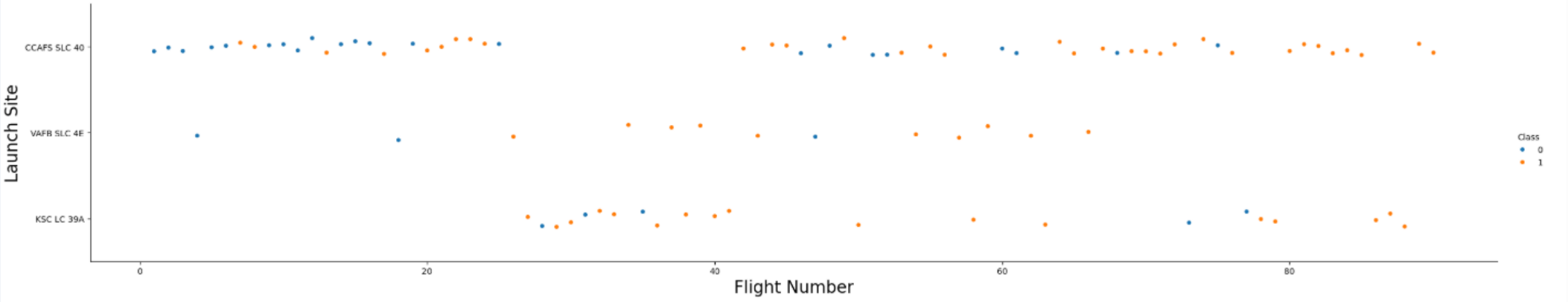
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

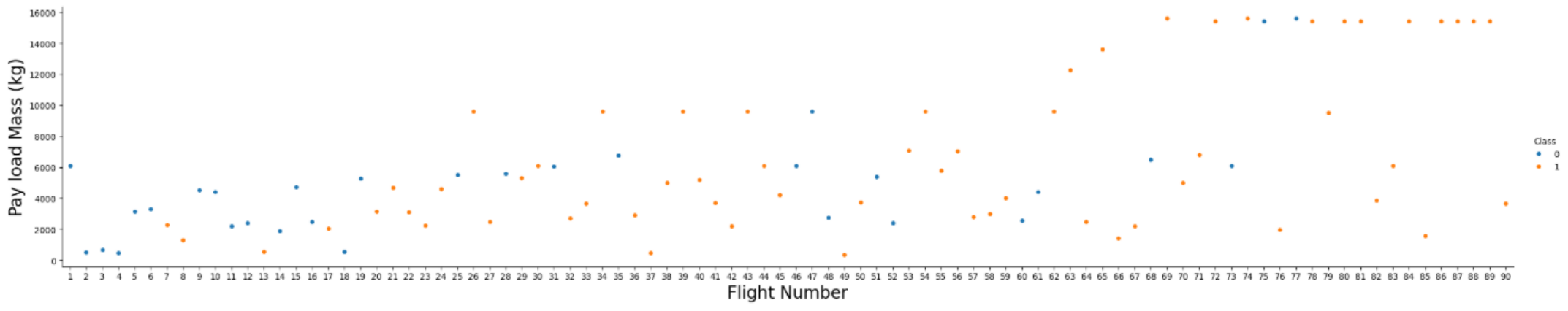
Insights drawn from EDA

Flight Number vs. Launch Site



- The number of launches from CCAFS SLC 40 were the highest.
- Falcon 9 launch from CCAFS SLC had achieved the greatest number of success.

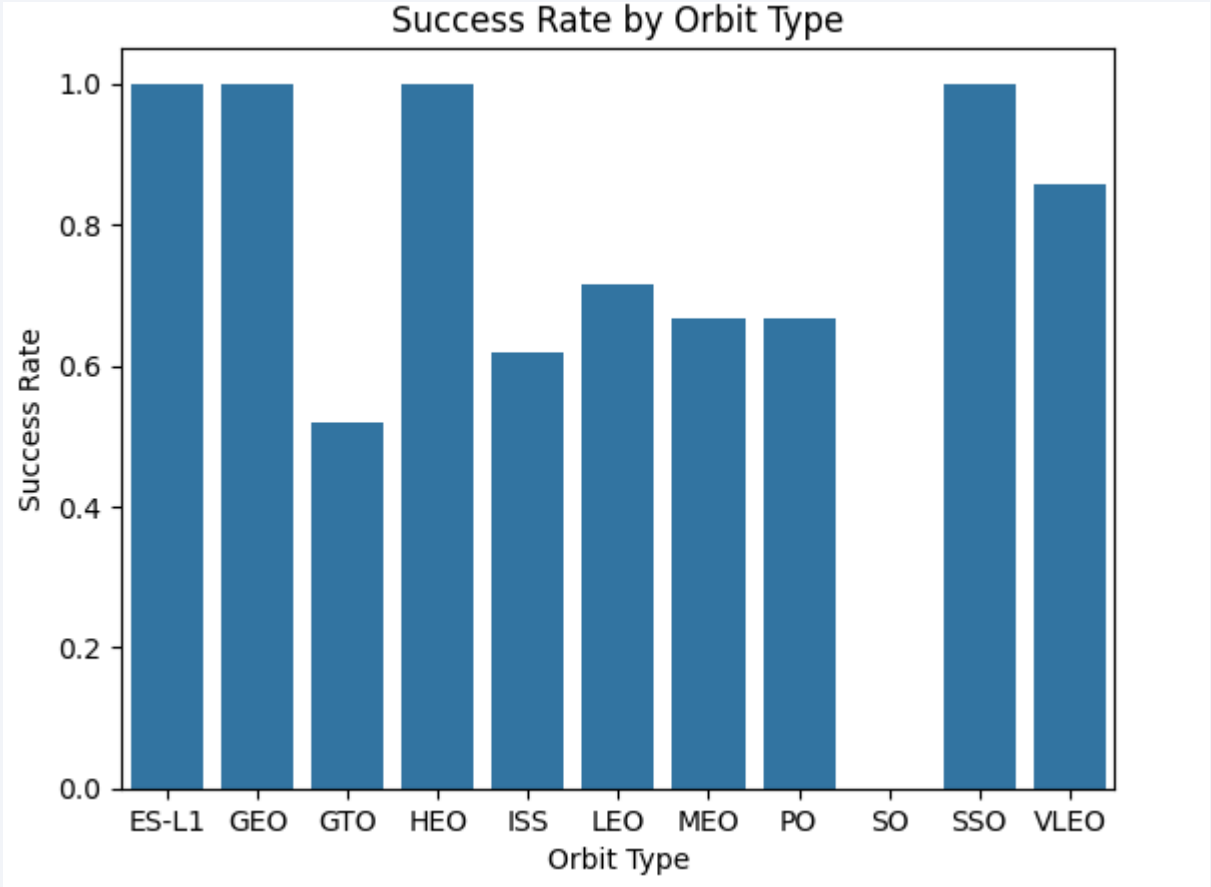
Payload vs. Launch Site



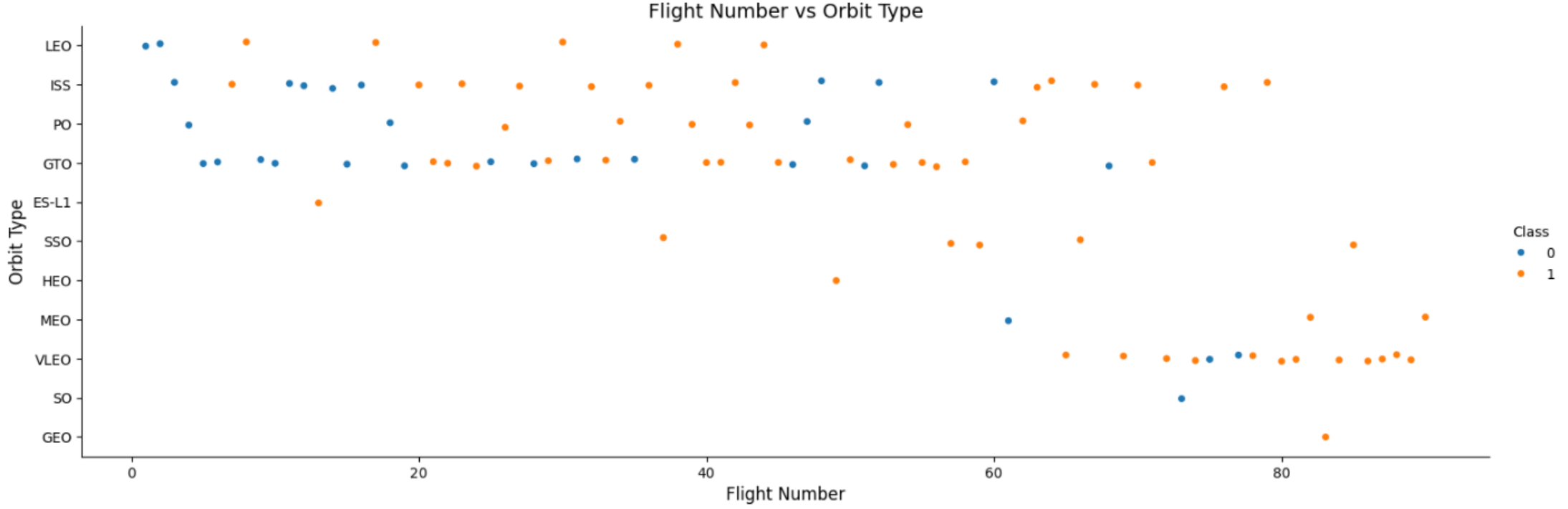
- It appears that the most success happened between a 4000 and 8000 kg payload.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have a 100% success rate.
- SO has zero success.

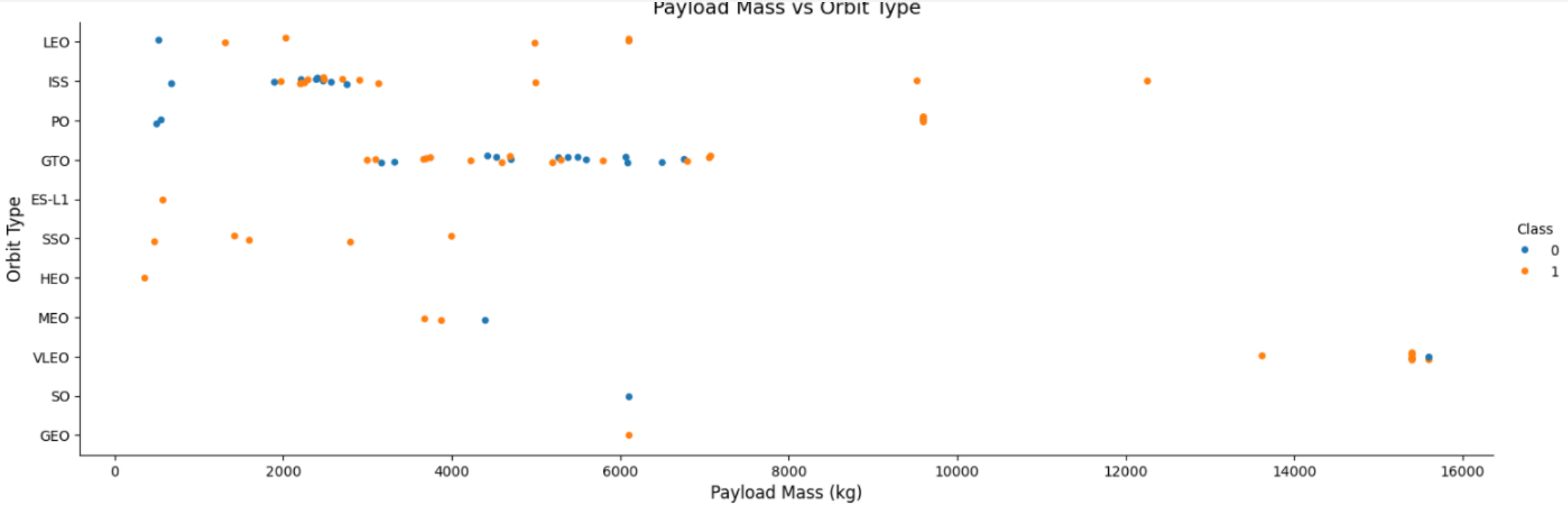


Flight Number vs. Orbit Type



- GTO appears to have the best success rate across the widest amount of flight numbers.

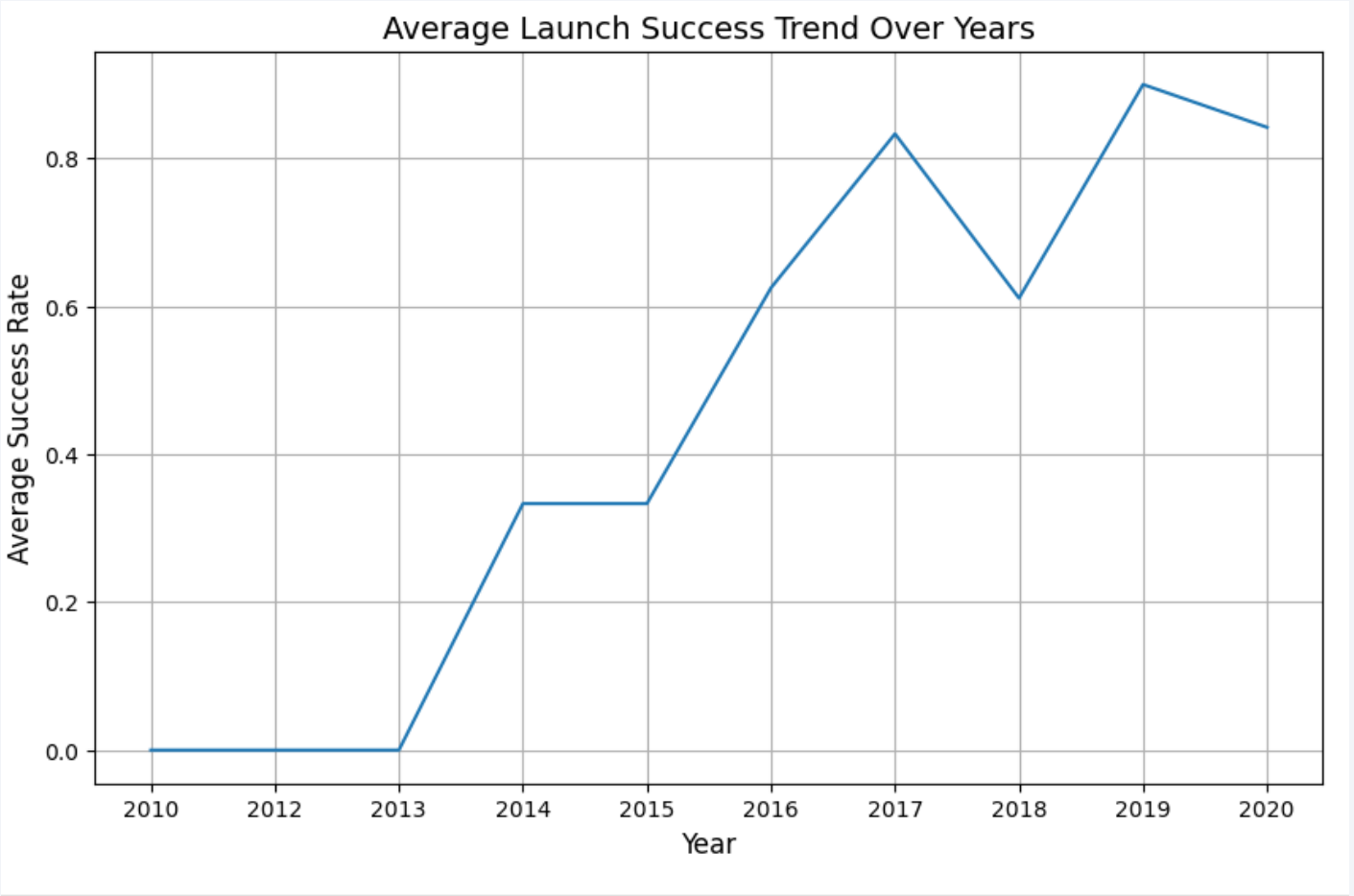
Payload vs. Orbit Type



- ISS seemed to do the best between 2,000 and 3,000 kg payload
- GTO had a span between 3,000 and 7,000 but seemed to have the best results between 5,000 and 6,500.

Launch Success Yearly Trend

- Success rate went over 60% after 2016 with the best success happening in 2019.



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.561857	0

- List of 5 launches where the Launch Site starts with CCA.

Total Payload Mass

- Total Payload Mass for NASA: 107,010

Average Payload Mass by F9 v1.1

- Average payload for F9 v1.1 is 2,928.4

First Successful Ground Landing Date

```
%sql select * FROM SPACEXTABLE ORDER BY Date, [Time (UTC)] LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)

Successful Boosters with Payload between 4000 and 6000

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

- Boosters with success on a payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Total
Failure (in flight)	1
Success	100

Boosters Carried Maximum Payload

Booster_Version	max_PAYLOAD_MASS__KG_
F9 B5 B1060.3	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1056.4	15600
F9 B5 B1051.6	15600
F9 B5 B1051.4	15600
F9 B5 B1051.3	15600
F9 B5 B1049.7	15600
F9 B5 B1049.5	15600
F9 B5 B1049.4	15600
F9 B5 B1048.5	15600
F9 B5 B1048.4	15600

2015 Launch Records

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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

Landing_Outcome	Count
Failure (drone ship)	5
Success (ground pad)	3

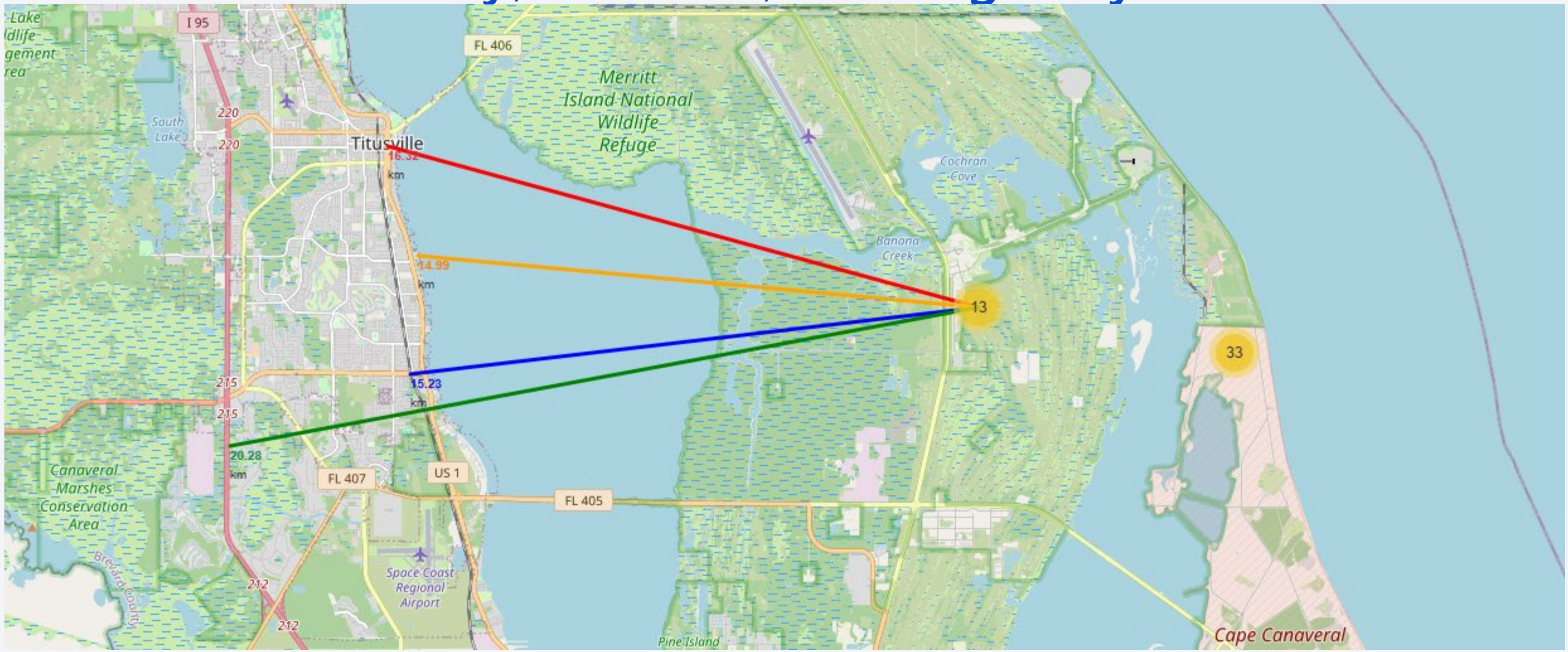
Section 3

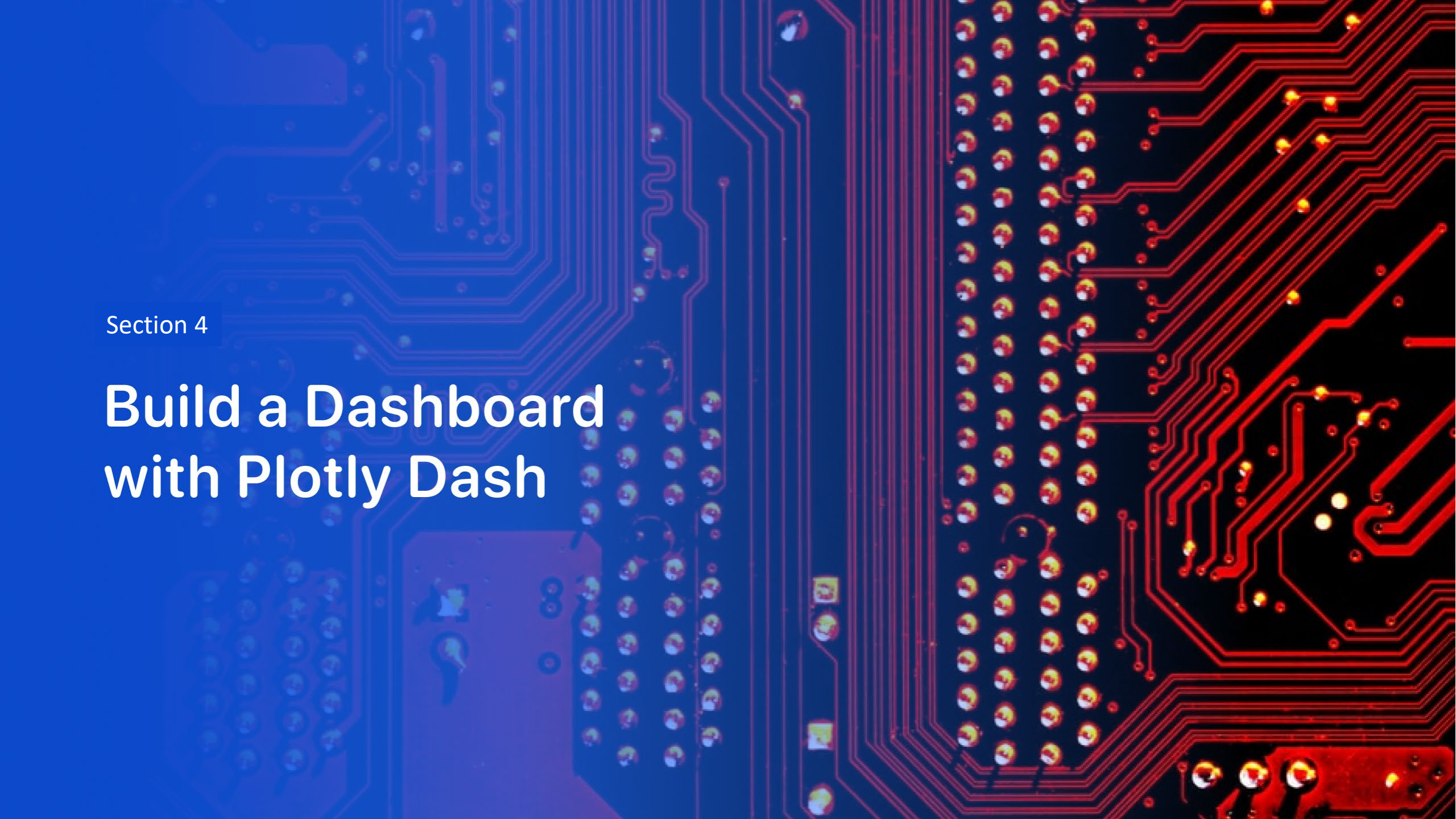
Launch Sites Proximities Analysis

Launch Sites

Launch Outcomes

Distance to City, Railroad, and Highway

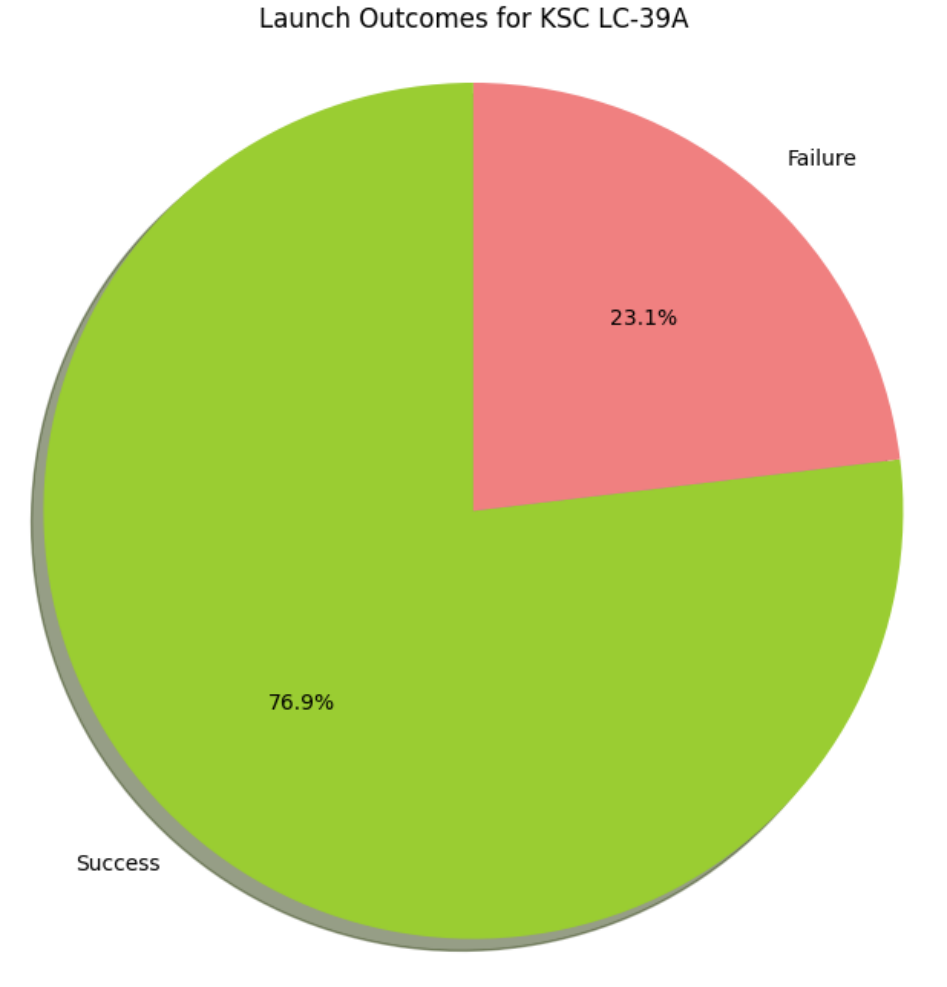



Section 4
Build a Dashboard with Plotly Dash

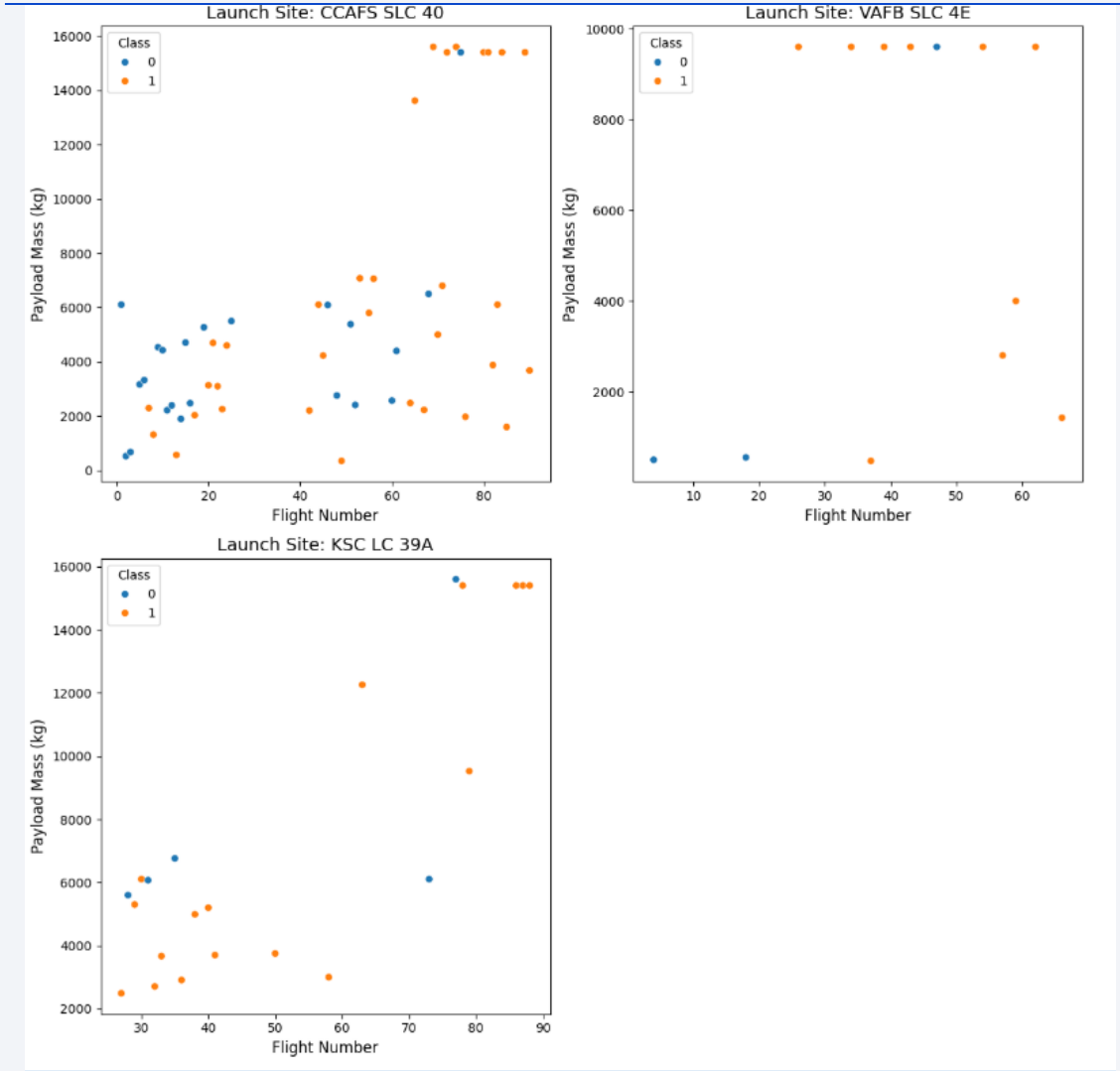
Launch Outcomes by Site

- KSC LC-39A is the most successful launch site.

Launch Outcomes for KSC LC-39A



Payload vs Outcomes by flight number

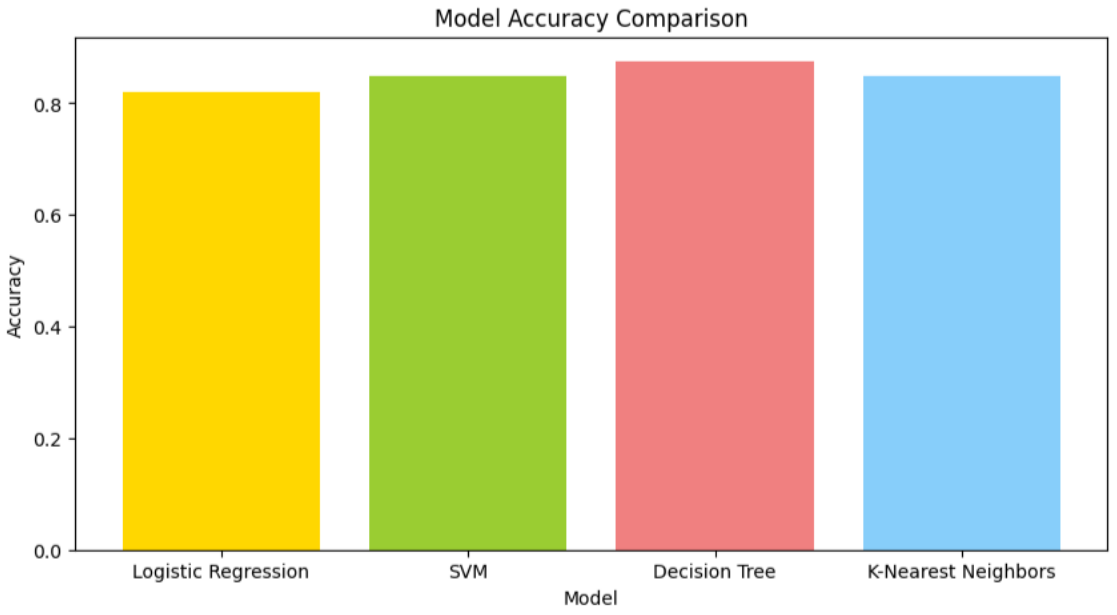


- CCAFS SLC 40 has the most success between 0 to 6000 payload.

Section 5

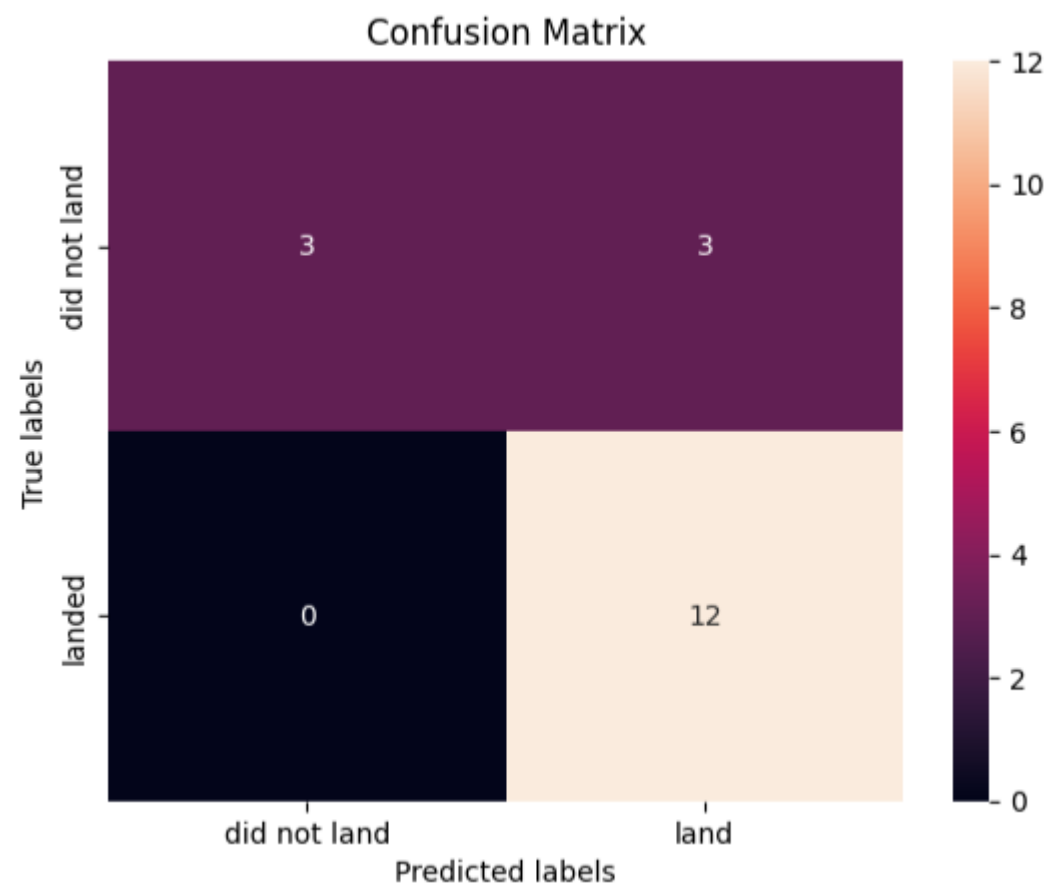
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree has the best accuracy.

Confusion Matrix



- Decision Tree has a low number of false positives.

Conclusions

In conclusion, this project demonstrated the feasibility of predicting Falcon 9 first stage landings using machine learning techniques. Logistic Regression and Decision Tree models achieved the highest accuracy, both at 87.5%. Exploratory data analysis highlighted the importance of factors such as flight number and payload mass in determining landing success. Accurate landing predictions can aid in cost estimation and risk assessment for space missions, ultimately impacting the economic viability of space exploration. While this study provides valuable insights, future work should focus on incorporating additional data sources, exploring advanced modeling techniques, and expanding the scope to other space-related applications

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!