Winning Space Race with Data Science

Julie Lovett
11/01/2025

Outline

- Executive Summary

- Introduction

- Methodology

- Results

- Conclusion

- Appendix

Executive Summary

- Summary of methodologies

  - Data was initially collected using multiple methods. The dataset was then wrangled to assess its structure and identify potential correlating factors. Exploratory data analysis (EDA) was conducted using SQL and Python to visualize and compare variables. Interactive visuals were developed to map and annotate geographic locations, enabling spatial calculations. Finally, predictive analysis was performed using classification models to evaluate and compare their accuracy.

- Summary of all results

  - The analysis revealed several key insights across SpaceX launch data. Among the four evaluated launch sites, CCAFS SLC-40 stood out with the highest success ratio at 42.9%. Booster version FT demonstrated the strongest performance for payloads between 2,000 and 4,000 kg, while F9 v1.1 averaged 2,534.67 kg per launch. The maximum payload mass recorded was 15,600 kg. NASA-sponsored missions contributed a total of 45,596 kg in payload volume. The first successful ground pad landing occurred on December 22, 2015, marking a pivotal milestone in reusability. Classification modeling showed the Decision Tree algorithm as the most accurate, with a confusion matrix confirming 16 correct predictions and only one misclassification. In total, 71 missions were evaluated, with a mix of successes and failures—many of which were planned test recoveries. These findings offer a comprehensive view of launch performance, booster reliability, and predictive modeling accuracy.

Introduction

- This project explores SpaceX rocket launch data to uncover relationships between key variables and gain deeper insight into launch outcomes.

- The primary goal of this analysis is to predict the cost of each SpaceX launch by determining whether the first stage will be reused. SpaceX advertises Falcon 9 launches at $62 million, significantly lower than competitors charging upwards of $165 million— a cost advantage largely driven by first-stage reusability. Instead of relying on rocket science to assess landing success, this project uses public mission data and machine learning models to predict whether the first stage will land successfully. By doing so, we can infer the likely cost of a launch and support strategic planning for future missions..

Section 1

Methodology

Draft

Methodology

Executive Summary

- Data collection methodology:

  - Data was collected by using an API and the GET method to extract information on SpaceX data and by Web scrapping from Wikipedia to collect Falcon 9 historical launch records.

- Perform data wrangling

  - The data was analyzed to find and clean missing data, evaluate the type of data, launch sites, and landing outcomes of each mission.

- Perform exploratory data analysis (EDA) using visualization and SQL

  - EDA was performed using SQL to analyze data further and to evaluate payload, booster types, landing outcomes and locations.

  - Python was also used with matplotlib to visualize comparisons of relationships in the launch records.

6

Methodology

Executive Summary cont.

- Perform interactive visual analytics using Folium and Plotly Dash

  - Folium was used to mark all of the launch sites, mark the success and failed launches on the map and calculate the distance between the launch site and its proximities.

  - Plotly Dash was used to perform calculations on each launch site, evaluate mission outcomes, including evaluating the success around various Orbits.

- Perform predictive analysis using classification models

  - The data was standardized and then put into training and testing data (train_test_split) to evaluate various classification models to determine the confusion matrix to find out the accuracy of each one. The classification models used were Logistic Regression, Support Vector Machines (SVMs), Decision Tree and K Nearest Neighbors (KNN).

Data Collection
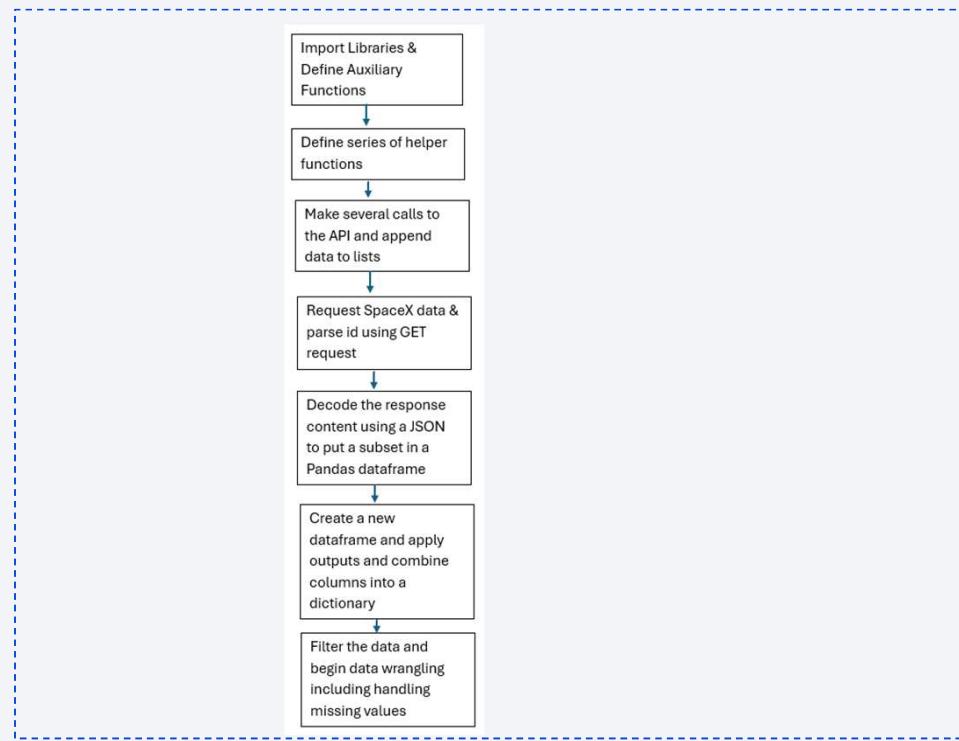
- Source: SpaceX launch dataset containing mission records from 2010 to early 2017.

- Scope: includes launch site names, booster versions, payload mass, customer affiliations, landing outcomes and mission dates.

- Filtering & Queries:

  - Extracted distinct launch sites and booster versions.

  - Filtered by payload mass ranges, landing outcomes, and date intervals.

  - Aggregated metrics such as total payload, average payload and success/failure counts.

- Purpose: to evaluate launch performance, booster reliability, and landing success patterns across SpaceX missions.

Data Collection – SpaceX API
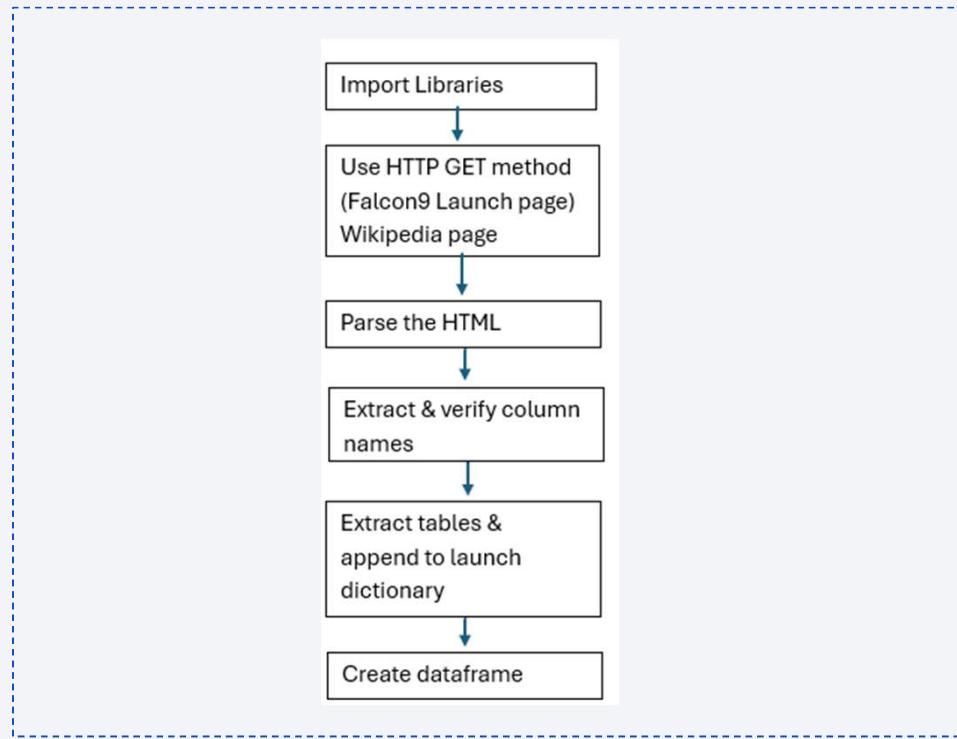
- Retrieved launch data via RESTful GET requests to SpaceX API. Parsed JSON responses and loaded relevant subsets into a Pandas dataframe. Created a new dataframe by combining selected columns into a structured dictionary and filtered and prepared the data for wrangling and downstream analysis.

- Capstone-Project/Data Collection SpaceXAPI.docx at main · PMDataGeek25/Capstone-Project

- Capstone-Project/API Flowchart.docx at main · PMDataGeek25/Capstone-Project

9

Data Collection - Scraping
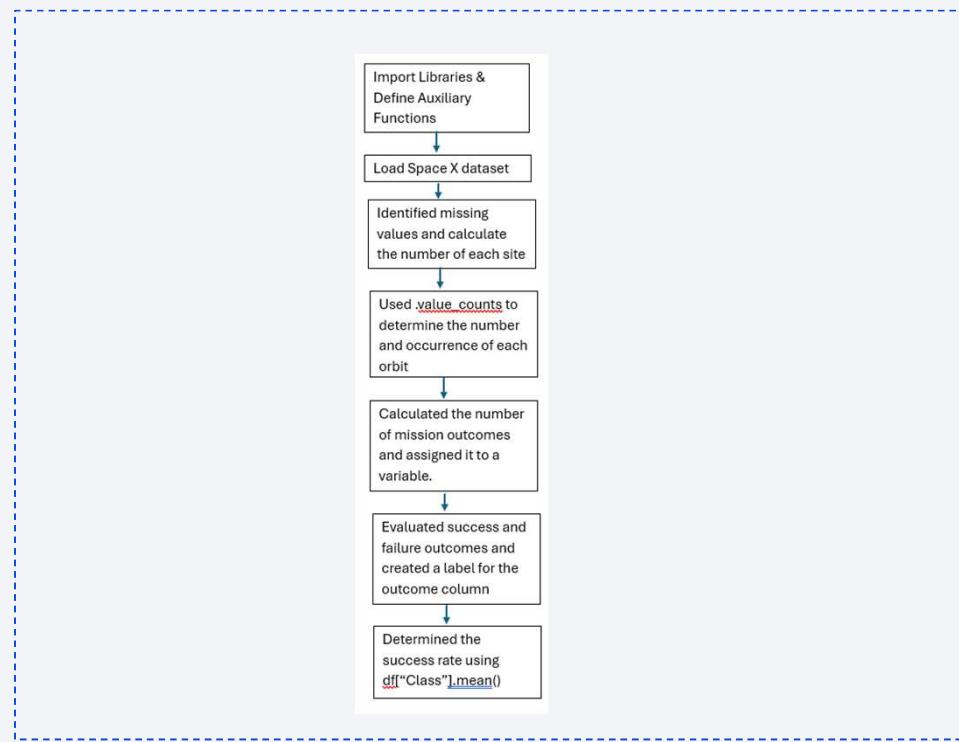
- Performed HTTP GET requests to extract table data from Wikipedia. Parsed HTML content and converted tables into structured dictionaries. Loaded the data into Pandas dataframes for further wrangling and analysis.

- Capstone-Project/Lab Webscrapping.docx at main · PMDataGeek25/Capstone-Project

- Capstone-Project/Webscrapping_flowchart.docx at main · PMDataGeek25/Capstone-Project

Import Libraries

Use HTTP GET method (Falcon9 Launch page) Wikipedia page

Parse the HTML

Extract & verify column names

Extract tables & append to launch dictionary

Create dataframe

Data Wrangling

- Conducted exploratory analysis to quantify launches per site. Aggregated launch counts to compare site activity. Encoded landing outcomes as binary values: Success (1), Failures (0). Calculated overall landing success rates for each site.

- Capstone-Project/Lab Data Wrangling.docx at main · PMDataGeek25/Capstone-Project

- Capstone-Project/Data Wrangling flowchart.docx at main · PMDataGeek25/Capstone-Project

11

EDA with Data Visualization

- Scatter plots visualized Flight Number, Payload Mass, and Launch site with flight outcomes overlaid. Used to assess where launch success increased with experience and to explore variable relationships.

- Bar chart analyzed success rates across Orbit Types. This was chose for its clarity in comparing categorical success rates.

- Scatter plot examined Payload Mass vs Orbit Type. Revealed which payloads were more likely to succeed in specific orbits.

- Line Graph tracked Launch Success Trends by Year. This highlighted temporal patterns and improvements over time.

- Capstone-Project/EDA with Data Visualization.docx at main · PMDataGeek25/Capstone-Project

12

EDA with SQL

- Calculated SUM and AVG of Payload Mass grouped by Customer and Booster Type.

- Used MIN to identify the first successful landing outcomes.

- Queried Booster Names with successful drone ship landings between 4000-6000 kg Payload Mass.

- Retrieved Booster Versions with MAX payload Mass.

- Extracted month names from 2015 to analyze temporal patterns across variables.

- Ranked landing outcomes between two dates in descending order to assess performance trends.

- Capstone-Project/EDA with SQL.docx at main · PMDataGeek25/Capstone-Project

Build an Interactive Map with Folium

Draft

- Plotted launch sites using latitude and longitude coordinates with circle markers and labeled popups.

- Visualized launch outcomes using color-coded markers and clusters to distinguish successes and failures.

- Calculated distances from each launch site to nearby features (e.g., coastlines, railroads, highways)

- Annotated the map with proximity markers to highlight spatial relationships and surrounding infrastructure.

- Capstone-Project/Interactive Map with Folium.docx at main · PMDataGeek25/Capstone-Project

14

Build a Dashboard with Plotly Dash

- Implemented a dropdown menu for dynamic Launch Site selection, allowing users to explore site-specific data.

- Created a pie chart with callback logic to display total launch outcomes (success vs failure) across all sites.

- Added a second pie chart to calculate and visualize success rates for each individual launch site.

- Integrated a slider with a scatter plot and callback to explore correlations between Payload Mass and launch success based on user input.

- Capstone-Project/Plotly Dash Lab.docx at main · PMDataGeek25/Capstone-Project

15

Predictive Analysis (Classification)

Draft
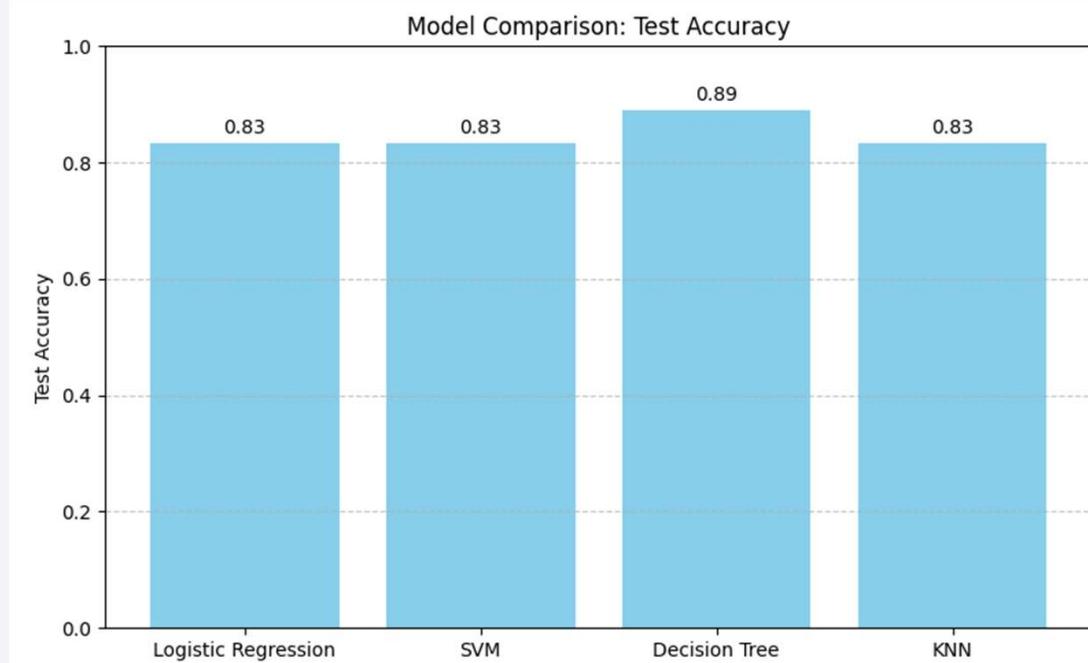
- Converted relevant features into a NumPy array for model input.

- Split the dataset into training and test sets to evaluate performance.

- Applied GridSearchCV to optimize hyperparameters across multiple classification models.

- Calculated accuracy scores to identify the best-performing model.

- Capstone-Project/Prediction Analysis.docx at main · PMDataGeek25/Capstone-Project

- Capstone-Project/Prediction flowchart.docx at main · PMDataGeek25/Capstone-Project

Results

- After evaluating multiple classification models, the Decision Tree emerged as the top performer.

- Achieved an accuracy of 89%, outperforming other models which averaged 83%

- The result highlights the Decision Tree's effectiveness in predicting launch outcomes based on available features.

- Capstone-Project/Prediction Analysis.docx at main · PMDataGeek25/Capstone-Project

17
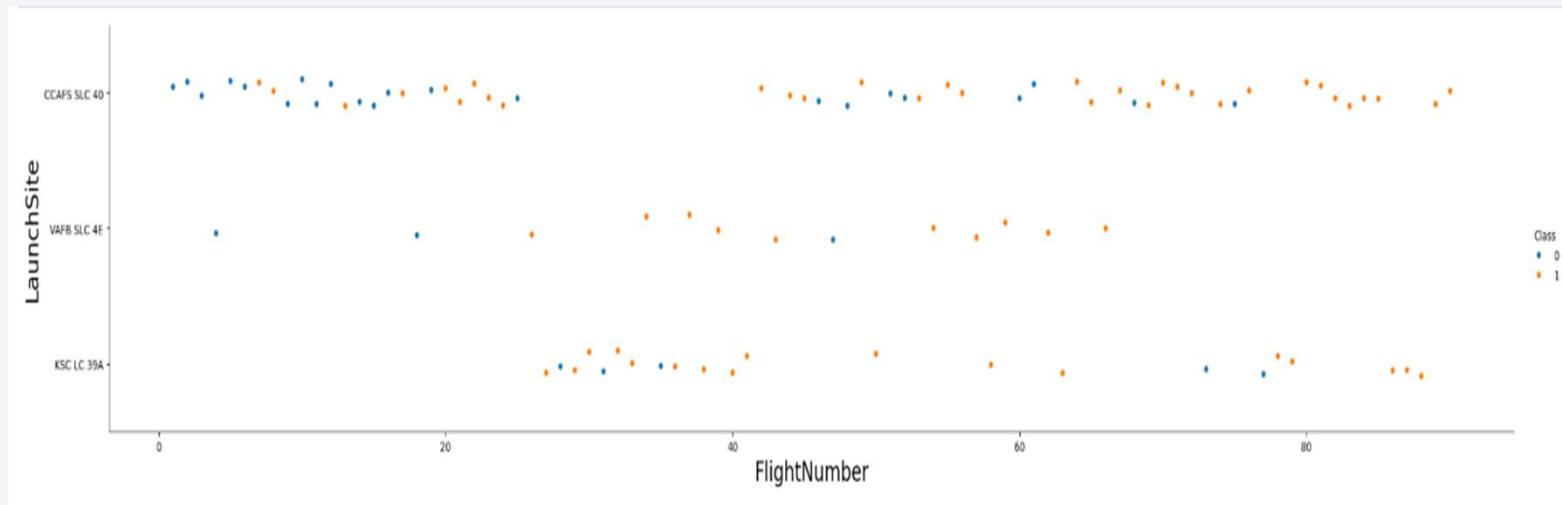
Draft

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site
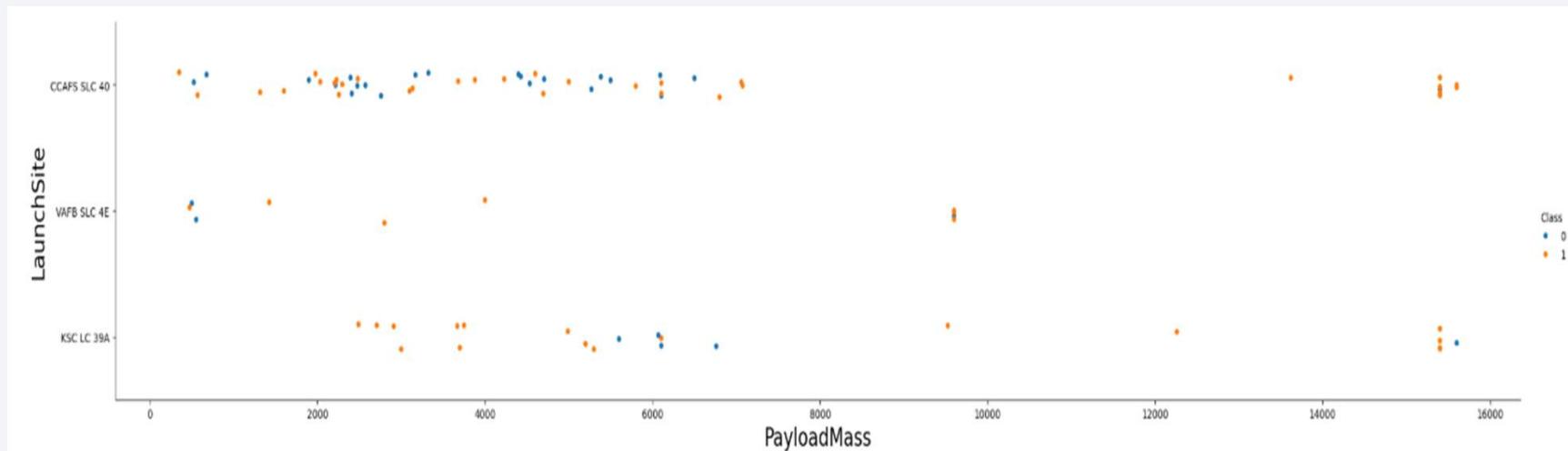
- This scatter plot illustrates the relationship between Flight Number and Launch Site, with launch outcomes labeled as success (1) or failure (0). The trend suggests that higher flight numbers are associated with increased launch success, indicating improved reliability over time.

Payload vs. Launch Site
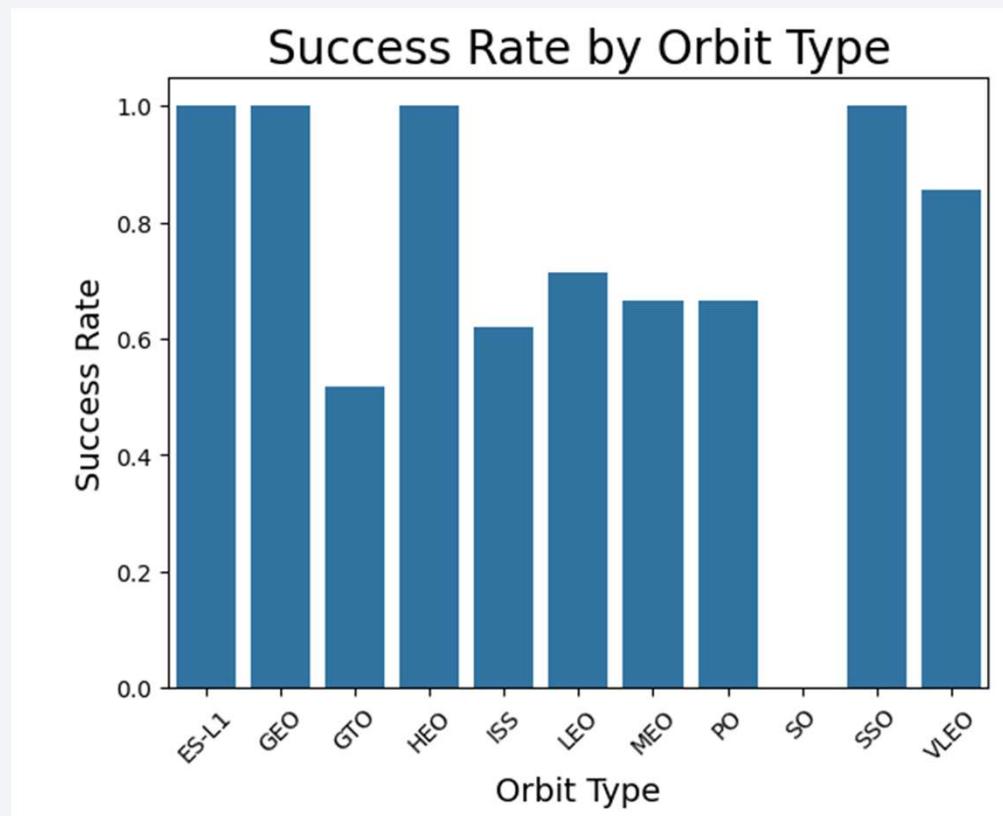
- This scatter plot compares the Payload Mass across different Launch Site, with launch outcomes indicated. The visualization reveals that not all sites achieve consistent success, but two locations show higher success rates when payloads exceed 14000 kg.

Success Rate vs. Orbit Type
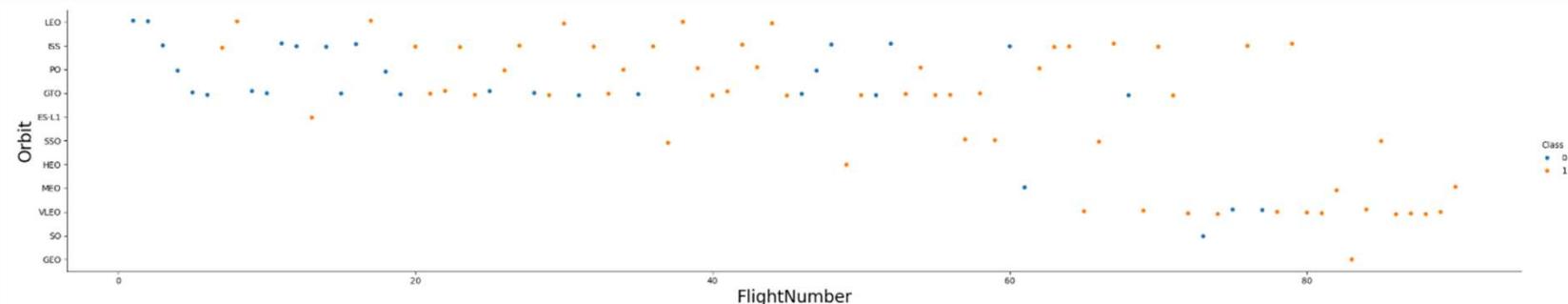
- This bar chart compares launch success rates across different Orbit Types. ES-L1, GEO, HEO, and SSO stand out with the highest success rates, indicating stronger reliability in these orbital missions.

Success Rate by Orbit Type

Flight Number vs. Orbit Type

Draft
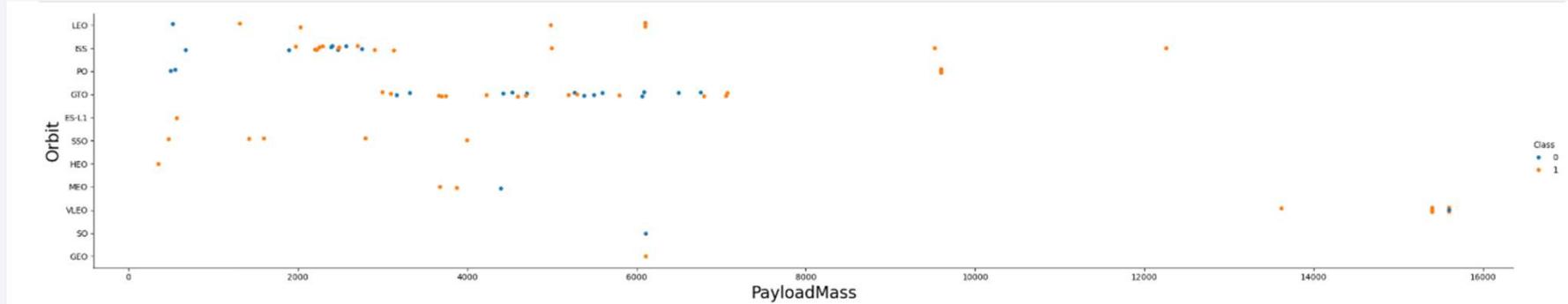
- This scatter plot highlights that launch success in LEO (Low Earth Orbit) increases with flight experience, suggesting a positive correctly with Flight Number. In contrast, GTO (Geostationary Transfer Orbit) shows no clear relationship. Notably, GTO serves as a transitional orbit and is not itself geostationary.

Payload vs. Orbit Type
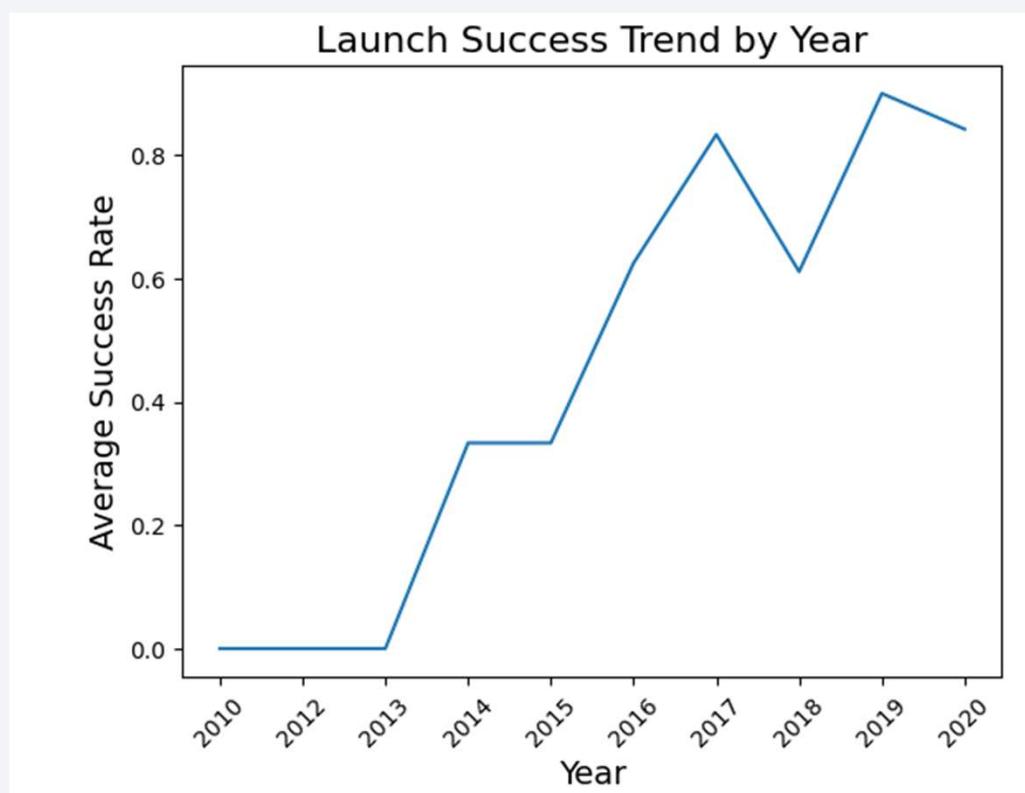
- This scatter plot compares Payload Mass across various Orbit Types, highlighting that Polar, LEO and ISS exhibit higher success rates. In contrast, GTO – a transitional orbit used for transfers - shows mixed outcomes, making direct comparisons more complex.

Launch Success Yearly Trend

- This line graph illustrates the yearly trend in launch success rates. A noticeable upward shift begins in 2013, with continued improvement in subsequent years – reflecting SpaceX's growing experience and refinement of launch strategies through iterative testing.

Launch Success Trend by Year

24

All Launch Site Names

- This query retrieved the distinct launch sites included in this evaluation. The sites are: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40. These represent the primary locations used for SpaceX launches during the dataset period.

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

Launch Site Names Begin with 'CCA'

Draft

- This query filtered launch site names that begin with 'CCA' returning 5 examples: CCAFS LC-40. These sites are part of the Cape Canaveral Air Force Station complex, frequently used for SpaceX launches.

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

Total Payload Mass

- This query calculated the total payload mass carried by boosters for NASA-sponsored missions, resulting in a combined payload of 45,596 kg. This highlights NASA's significant contribution to the overall launch volume in the dataset.

SUM(PAYLOAD_MASS__KG_)

45596

Average Payload Mass by F9 v1.1

- This query calculated the average payload mass for launches using the Booster Version F9 v1.1, resulting in an average of approximately 2534.67 kg. This provides insight into the typical payload capacity for this booster configuration.
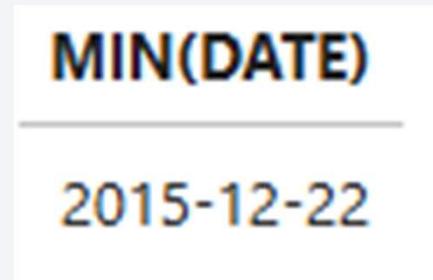
AVG(PAYLOAD_MASS__KG_)

2534.6666666666665

First Successful Ground Landing Date

- This query identified the first successful landing outcome on ground pad, which occurred on December 22, 2015. This milestone marked a major breakthrough in SpaceX's reusability efforts.

MIN(DATE)

2015-12-22

# Successful Drone Ship Landing with Payload between 4000 and 6000
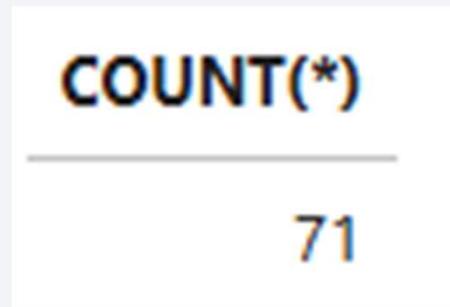
- This query retrieved the names of boosters that successfully landed on drone ship carried payloads between 4000 and 6000 kg. These results highlight mid-range payload missions with successful offshore recoveries.

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- This query calculated the total number of mission outcomes, combining both successful and failed launches. The result was 71 missions in total, providing a complete count of evaluated launch attempts.

COUNT(*)

71

Boosters Carried Maximum Payload

- This query identified the booster(s) that carried the maximum payload mass of 15,600 kg. These missions represent the upper limit of payload capacity in the dataset, showcasing the booster's peak performance.

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

2015 Launch Records

- This query retrieved failed landing outcomes in drone ships during the year 2015, along with the associated booster versions and launch site names. These results help identify which configurations and locations were involved in early offshore recovery attempts.
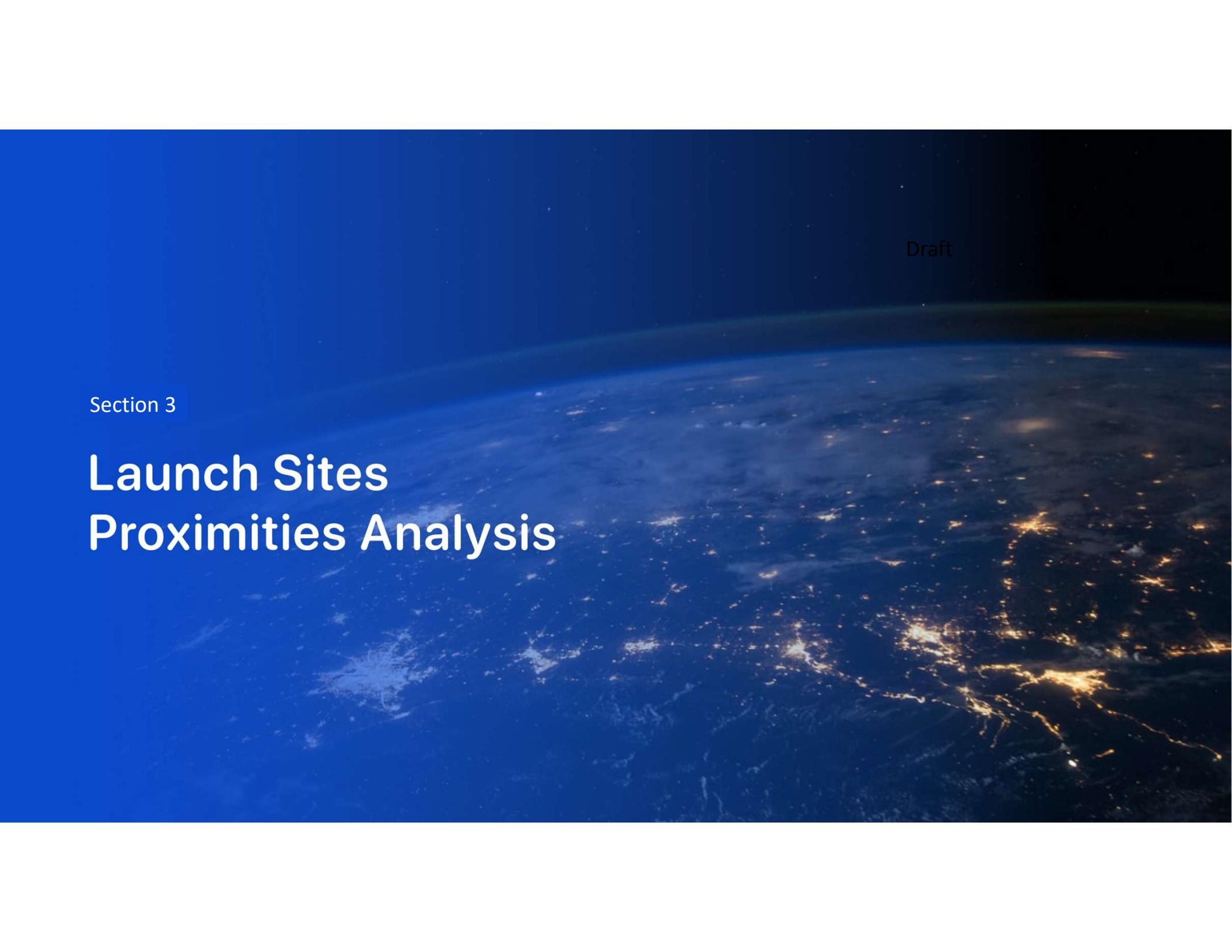
Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This query ranked landing outcome types – such as Failure (drone ship) and Success (ground pad) by their frequency between June 4, 2010, and March 20, 2017, sorted in descending order. The results highlight which outcomes were most common during SpaceX's early landing attempts.

Landing_Outcome	Outcome_Count	Rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

34
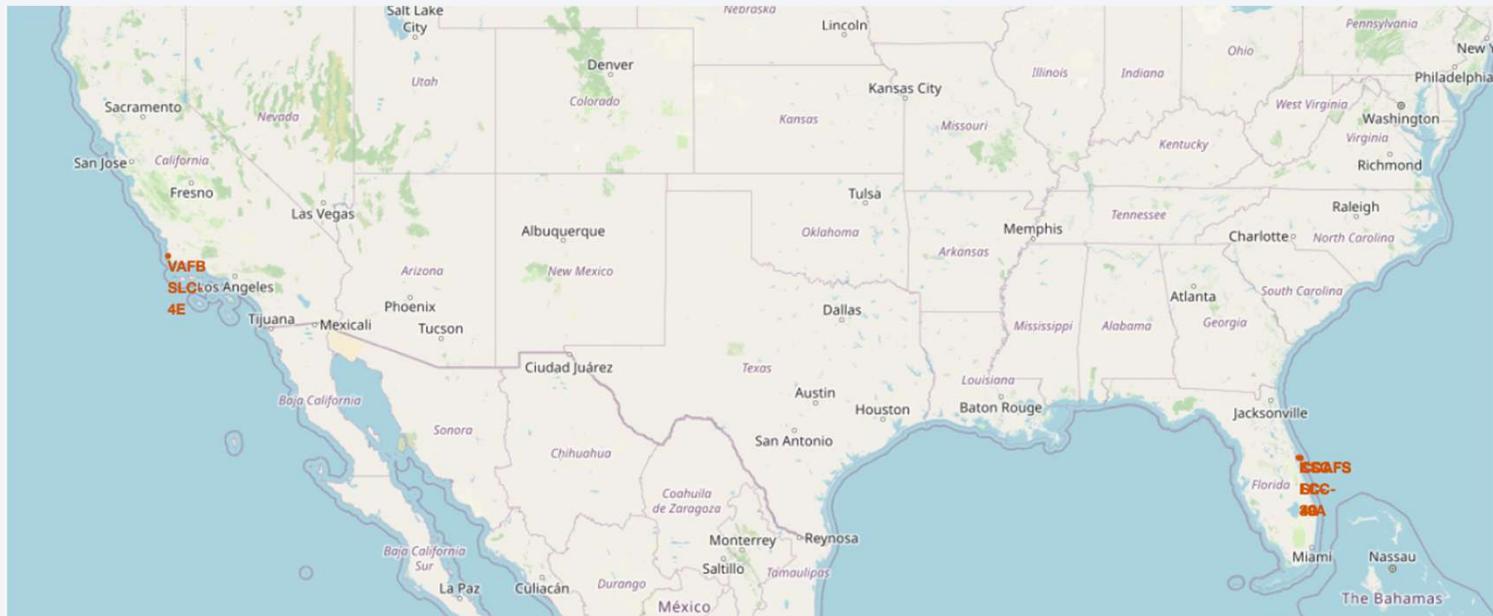
Draft

Section 3

Launch Sites
# Proximities Analysis

Launch sites with markers on a map

Draft
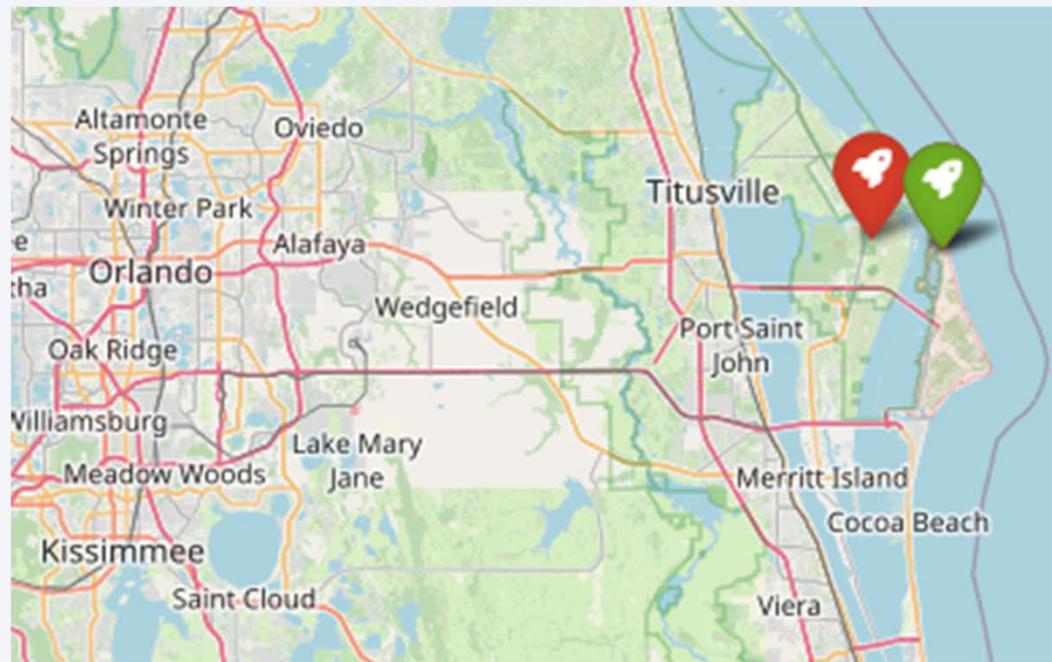
- Explain the important elements and findings on the screenshot

Launch Outcomes on a map

- This is for location CCAFS SLC-40 with the successful and failed launch outcomes on a map.

Distance between launch site and nearby proximities
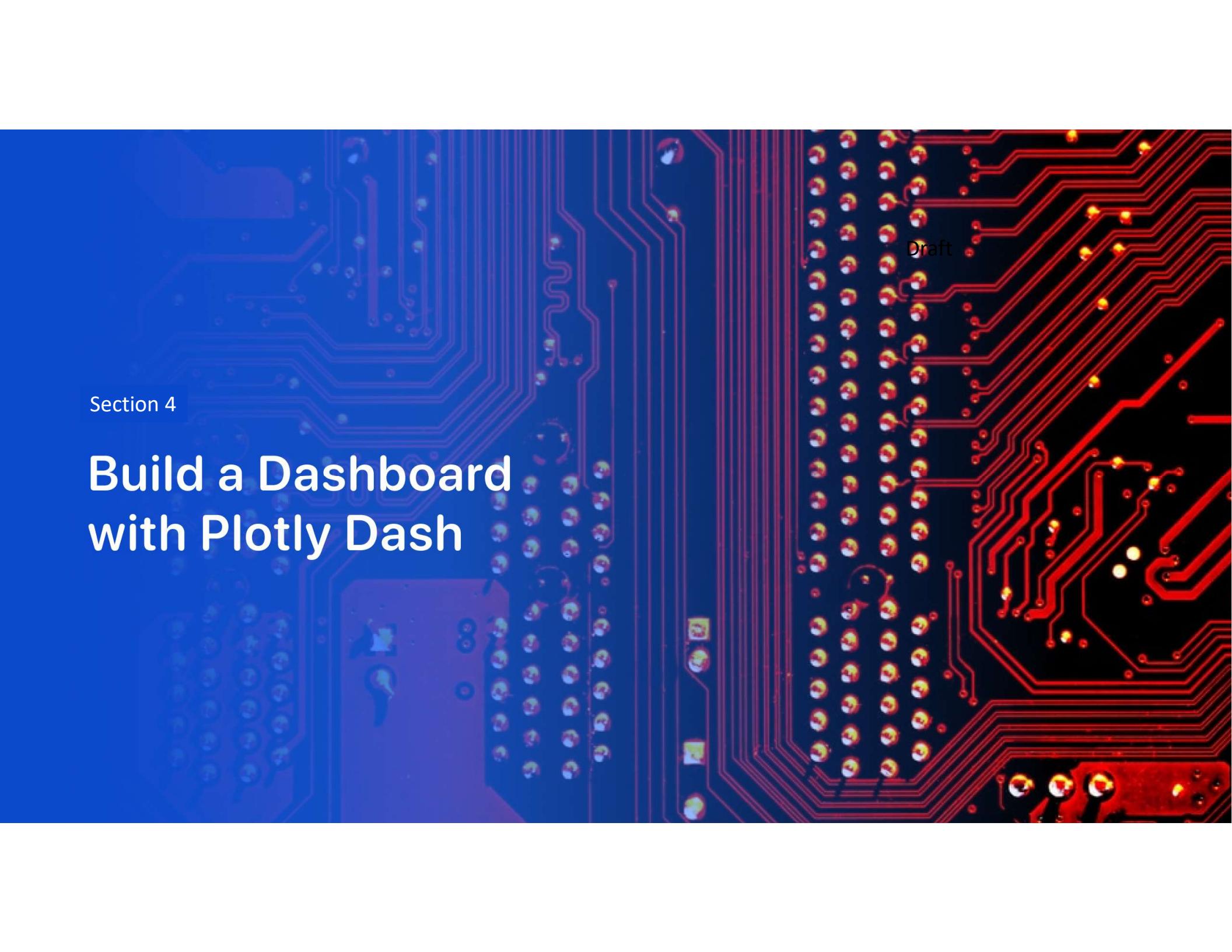
Draft

- My Folium map was not rending properly. I was able to determine the distance between CCAFS LC-40 and nearby proximities. Which shows that each of these launch sites are nearest to a Coastline and are relatively close to Railway's and Highway's.

```
Distance to Railway: 0.77 km
Distance to Highway: 0.89 km
Distance to Coastline: 0.74 km
```

38
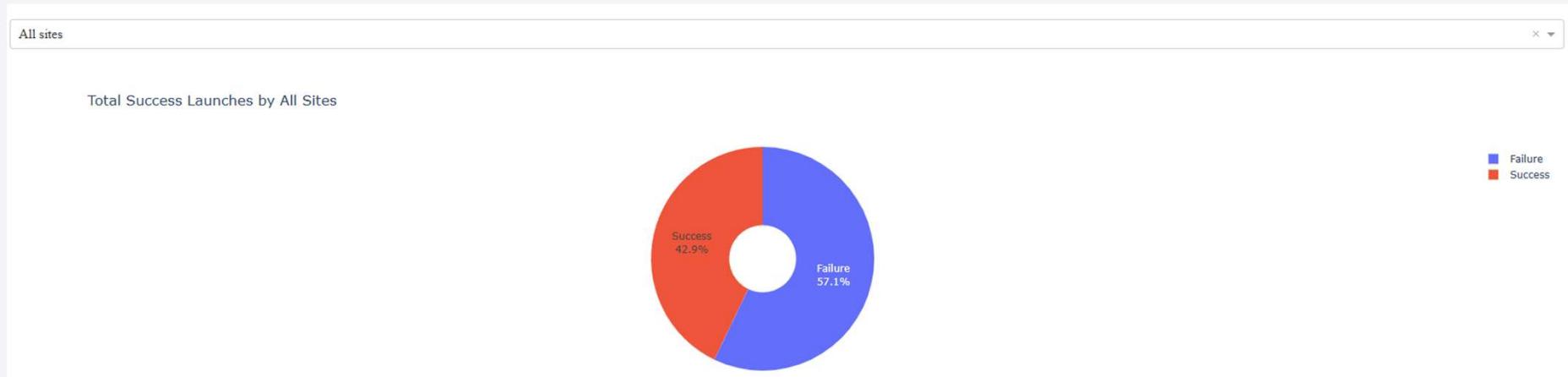
Section 4

Build a Dashboard
with Plotly Dash

Draft

# Total Success Launches by All site
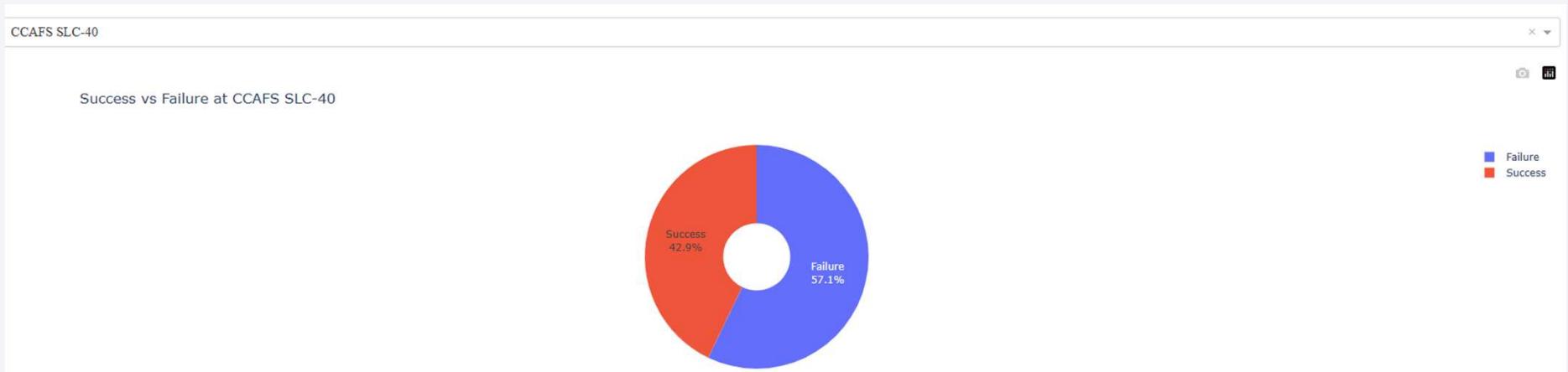
- This pie chart shows that at 42.9% of launches across all evaluated sites were successful, while 57.1% were classified as failures. However, it's important to note that many of these unsuccessful landings were intentional or plan, often part of test missions or low-priority recover attempts.

Launch Site with highest launch success ratio

- This pie chart highlights CCAFS SLC-40 as the launch site with the highest success ratio, accounting for 42.9% of all successful launches among the evaluated locations.

Booster Version with the largest Success Rate

Draft

- This scatter plot shows that the FT Booster Version achieved the highest success rates for payloads between 2,000 and 4,000 kg. This suggests that this configuration performs optimally within that payload range.
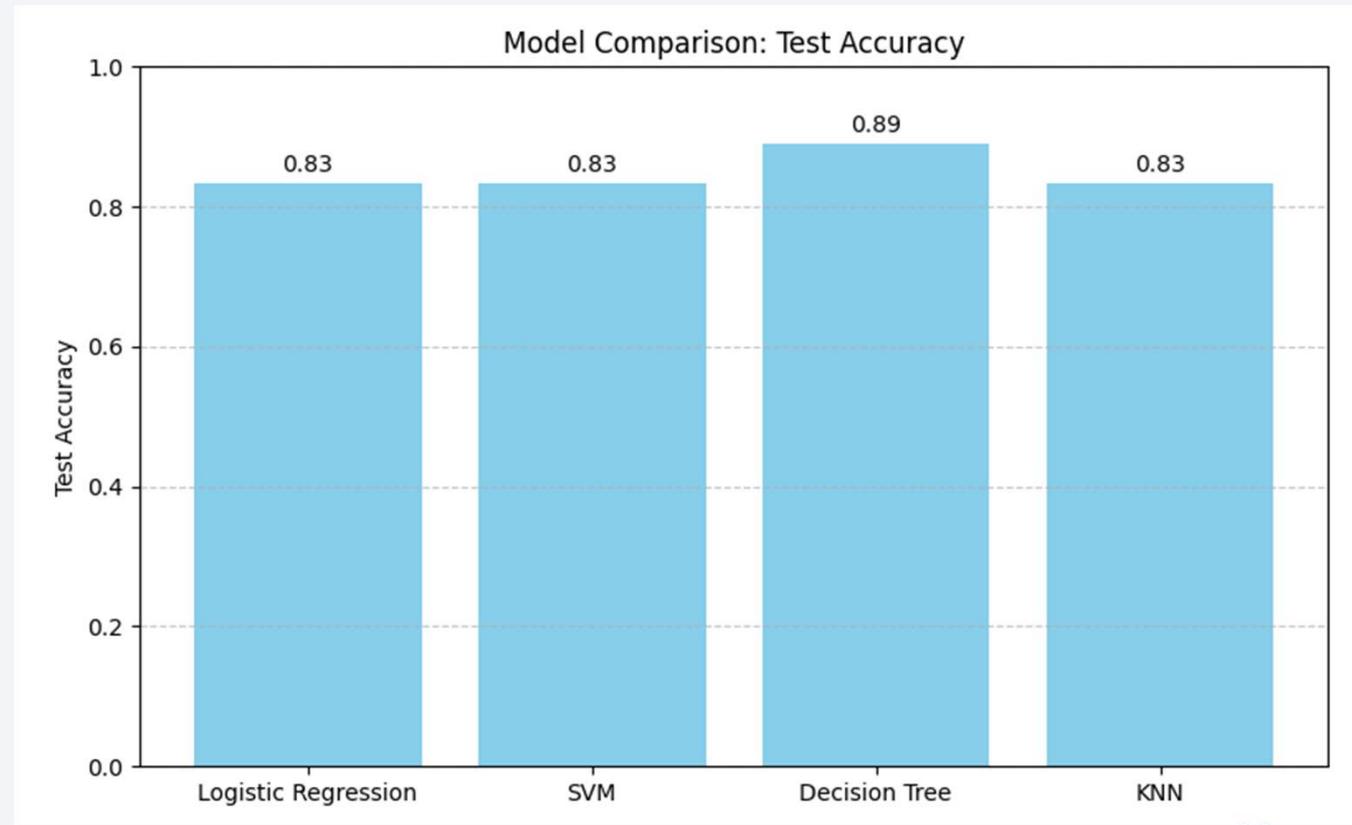
Draft

Section 5

Predictive Analysis (Classification)

Classification Accuracy
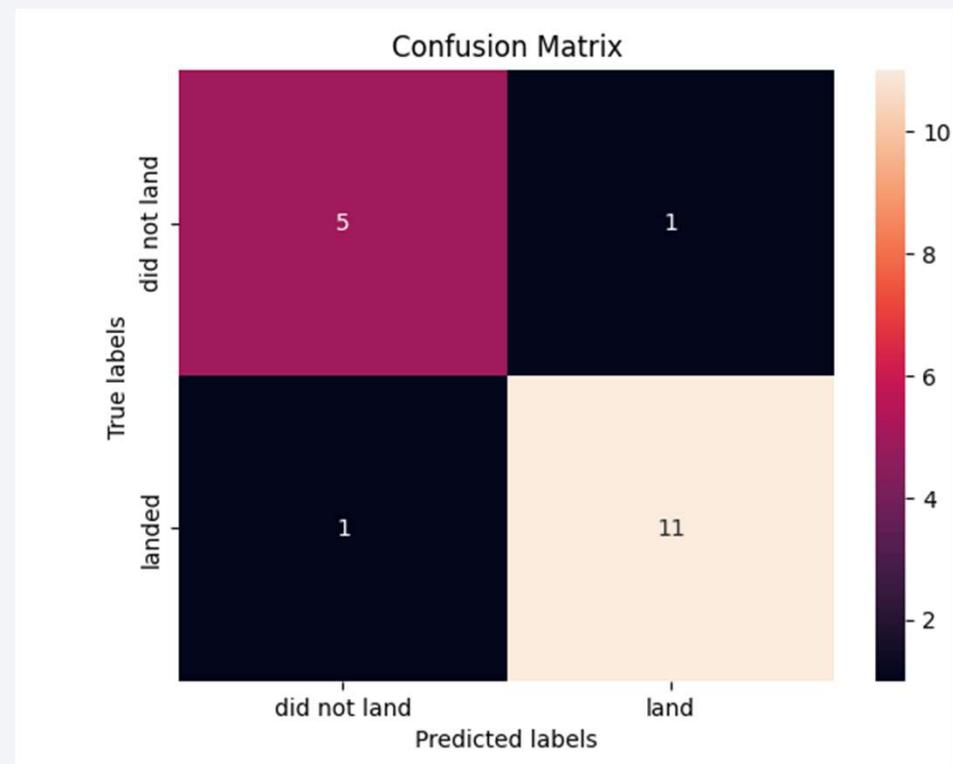
- Among all classification models evaluated, the Decision Tree achieved the highest accuracy, outperforming other approaches in predicting launch outcomes.

Model Comparison: Test Accuracy

Confusion Matrix

- This Confusion Matrix shows that the Decision Tree model was highly accurate, correctly predicting 5 non-landings and 11 successful landings, with only 1 misclassification between the two outcomes. This reinforces the model's reliability in distinguishing launch success.

45

Conclusions

- Booster FT demonstrated optimal reliability for mid-weight payloads.

- CCAFS SLC-40 emerged as the most successful launch site.

- Decision Tree model outperformed other classifiers in predicting landing outcomes.

- Confusion Matrix confirmed strong model accuracy with minimal misclassification.

- Landing success milestones began in late 2015, marking a turning point in reusability.

- NASA missions contributed significantly to total payload volume.

- Data-driven insights support strategic planning for future launch configurations and recovery methods.

46

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

Draft

Thank you!