Winning Space Race with Data Science

Julie Lovett
11/01/2025

Outline

Draft

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

Executive Summary

Draft

- 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

Draft

- 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

Methodology

Draft

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.

Methodology

Draft

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

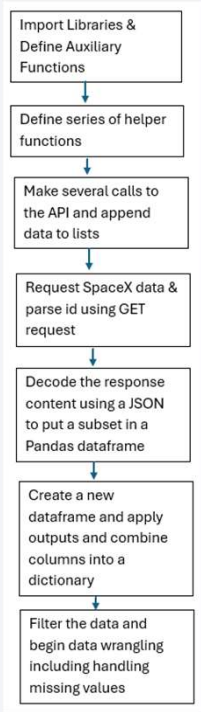
Draft

- 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

Draft

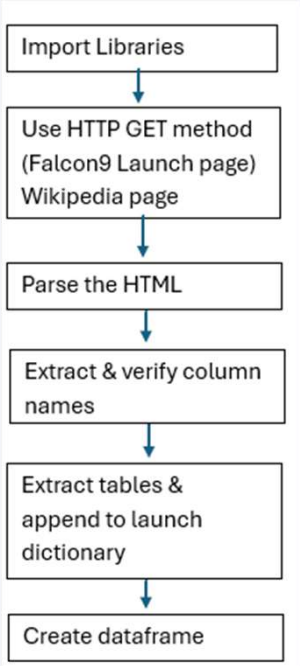
- Retrieved launch data via RESTful GET requests to SpaceX API. Parsed JSON responses and loaded relevant subsets into a Pandas dataframe. Created a new dataframe by combining selected columns into a structured dictionary and filtered and prepared the data for wrangling and downstream analysis.
- Capstone-Project/Data Collection SpaceXAPI.docx at main · PMDataGeek25/Capstone-Project
- Capstone-Project/API Flowchart.docx at main · PMDataGeek25/Capstone-Project



Data Collection - Scraping

Draft

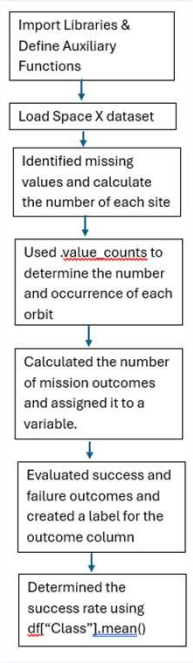
- Performed HTTP GET requests to extract table data from Wikipedia. Parsed HTML content and converted tables into structured dictionaries. Loaded the data into Pandas dataframes for further wrangling and analysis.
- Capstone-Project/Lab Webscrapping.docx at main · PMDataGeek25/Capstone-Project
- Capstone-Project/Webscrapping_flowchart.docx at main · PMDataGeek25/Capstone-Project



Data Wrangling

Draft

- 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



EDA with Data Visualization

Draft

- 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

EDA with SQL

Draft

- 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

Build a Dashboard with Plotly Dash

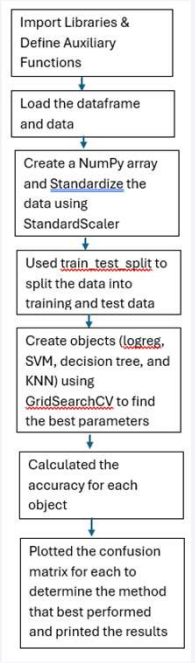
Draft

- Dynamic Launch site Selection via Dropdown Menu
  - Developed an intuitive dropdown interface enabling users to filter and explore launch data by site. This feature empowers stakeholders to analyze site-specific performance metrics and trends with ease, enhancing data interactively and decision-making.
- Aggregate Launch Outcome Visualization (Pie Chart)
  - Implemented a responsive pie chart powered by callback logic to display the overall distribution of launch outcomes – success vs. failure – across all sites. The high-level view provides immediate insight into mission reliability and operational consistency.
- Site-Level Success Rate Breakdown (Second Pie chart)
  - Added a secondary pie chart that dynamically calculates and visualizes the success rate for each selected launch site. This granular perspective supports comparative analysis and highlights performance variability across locations.
- Payload Mass vs Launch Success Correlation (Slider + Scatter Plot)
  - Integrated a user-controlled slider to filter payload mass ranges, triggering a scatter plot update via callback. This visualization reveals potential correlations between payload weight and launch success, offering actionable insights into mission planning and payload optimiazation.
- Capstone-Project/Plotly Dash Lab.docx at main · PMDataGeek25/Capstone-Project

Predictive Analysis (Classification)

Draft

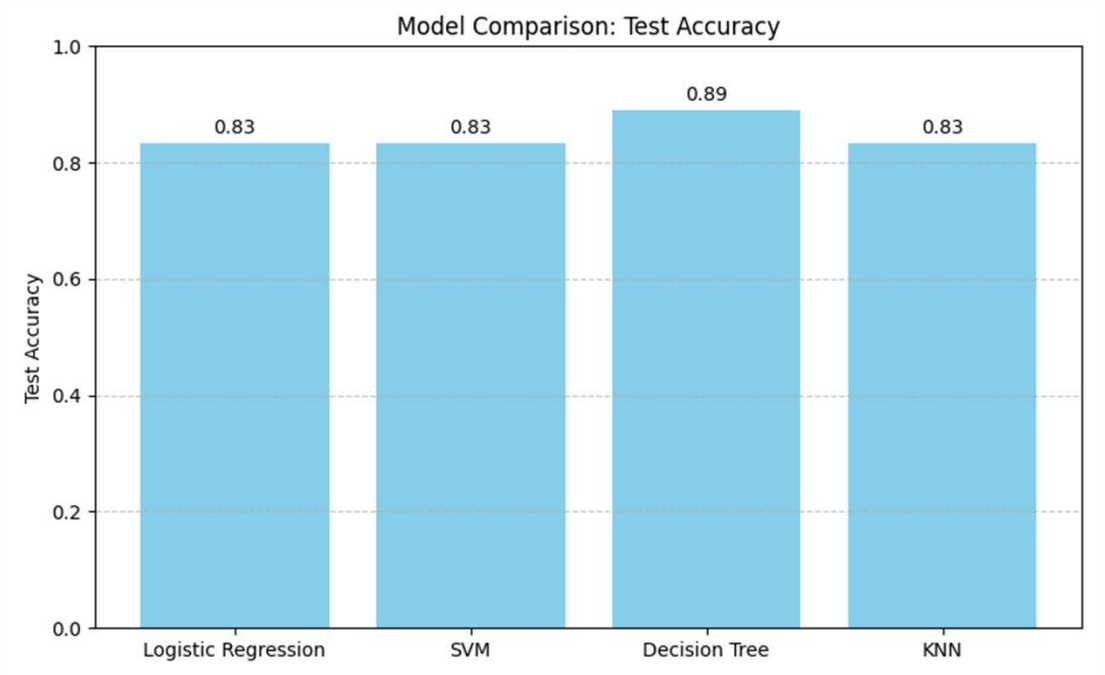
- Converted relevant features into a NumPy array to prepare structured input for machine learning models.
- Split the dataset into training and test sets to evaluate model performance and generalization.
- Applied GridSearchCV to optimize hyperparameters across multiple classification algorithms, including Logistic Regression, SVM, and Decision Trees.
- Calculated accuracy scores to identify the best-performing model and assess predictability.
- Capstone-Project/Prediction Analysis.docx at main · PMDataGeek25/Capstone-Project
- Capstone-Project/Prediction flowchart.docx at main · PMDataGeek25/Capstone-Project



Results

Draft

- 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]



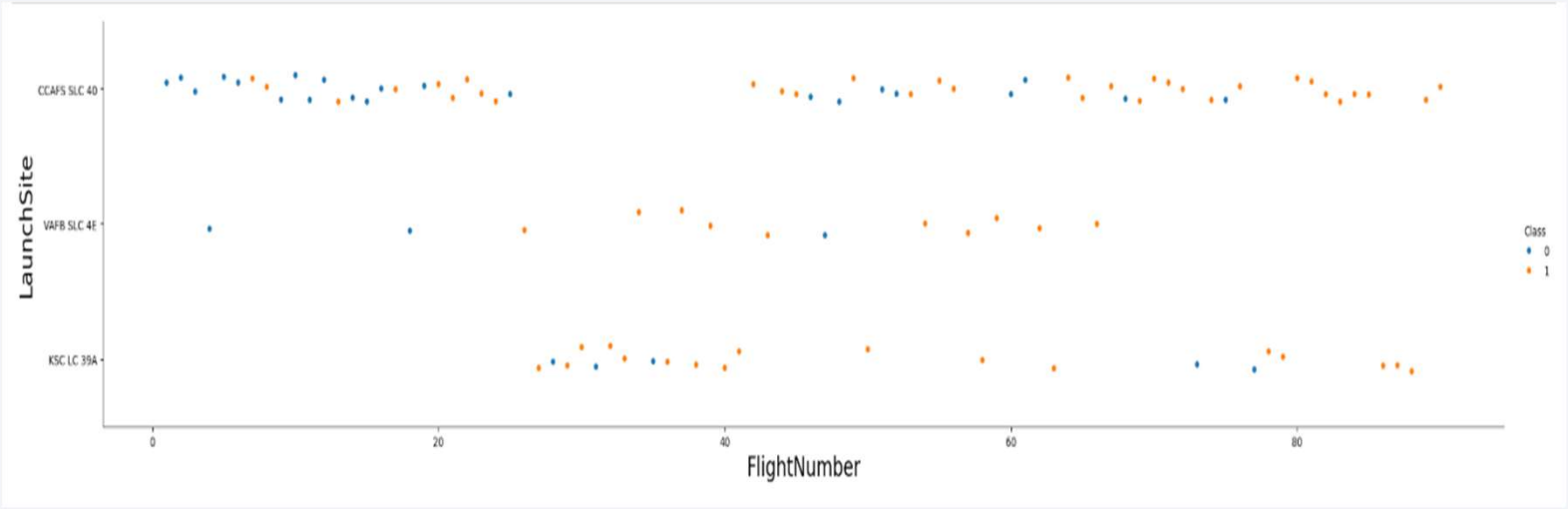
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Draft

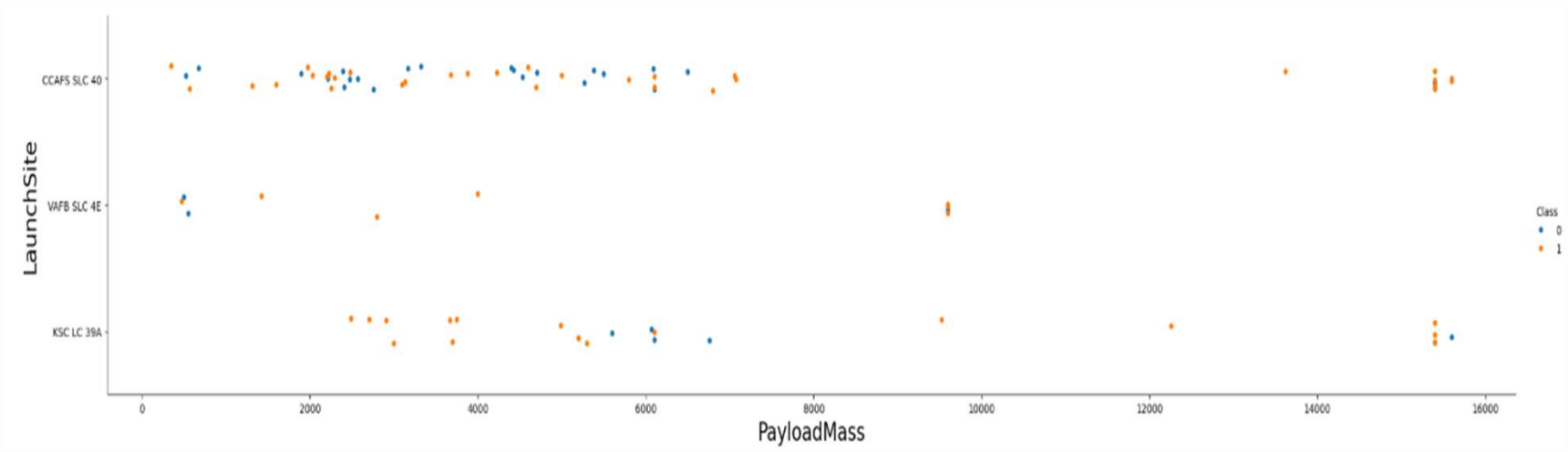
- This scatter plot illustrates the relationship between Flight Number and Launch Site, with launch outcomes labeled as success (1) or failure (0). The trend suggests that higher flight numbers are associated with increased launch success, indicating improved reliability over time.



Payload vs. Launch Site

Draft

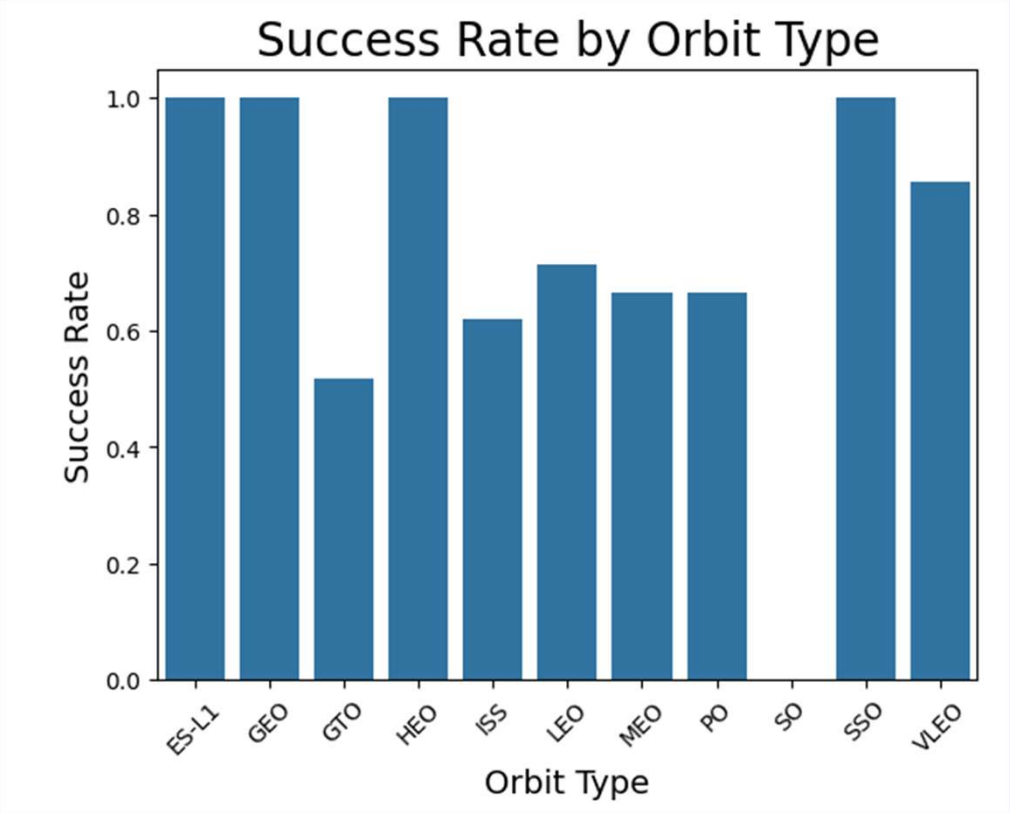
- This scatter plot compares the Payload Mass across different Launch Site, with launch outcomes indicated. The visualization reveals that not all sites achieve consistent success, but two locations show higher success rates when payloads exceed 14000 kg.



Success Rate vs. Orbit Type

Draft

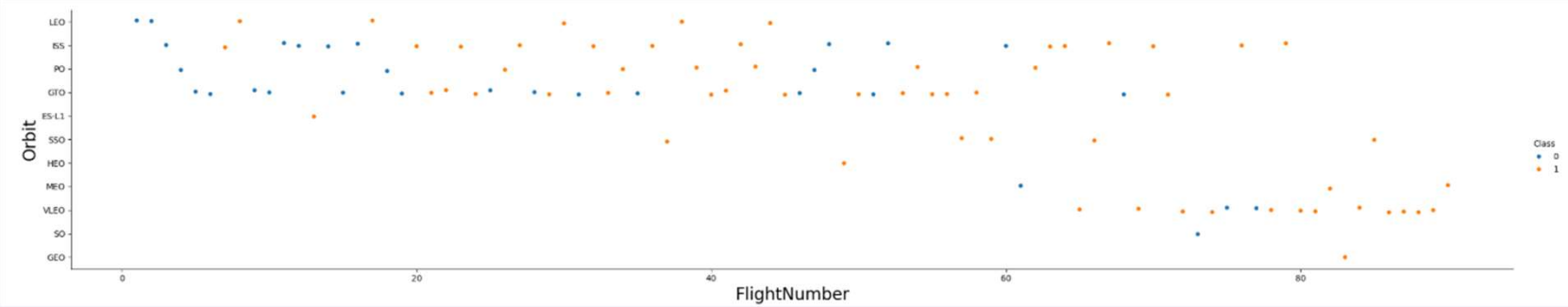
- This bar chart compares launch success rates across different Orbit Types. ES-L1, GEO, HEO, and SSO stand out with the highest success rates, indicating stronger reliability in these orbital missions.



Flight Number vs. Orbit Type

Draft

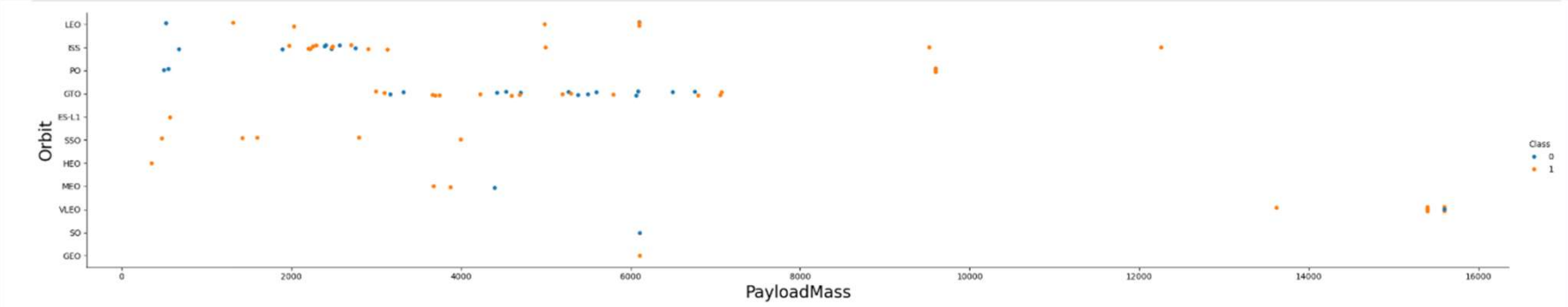
- This scatter plot highlights that launch success in LEO (Low Earth Orbit) increases with flight experience, suggesting a positive correctly with Flight Number. In contrast, GTO (Geostationary Transfer Orbit) shows no clear relationship. Notably, GTO serves as a transitional orbit and is not itself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



Payload vs. Orbit Type

Draft

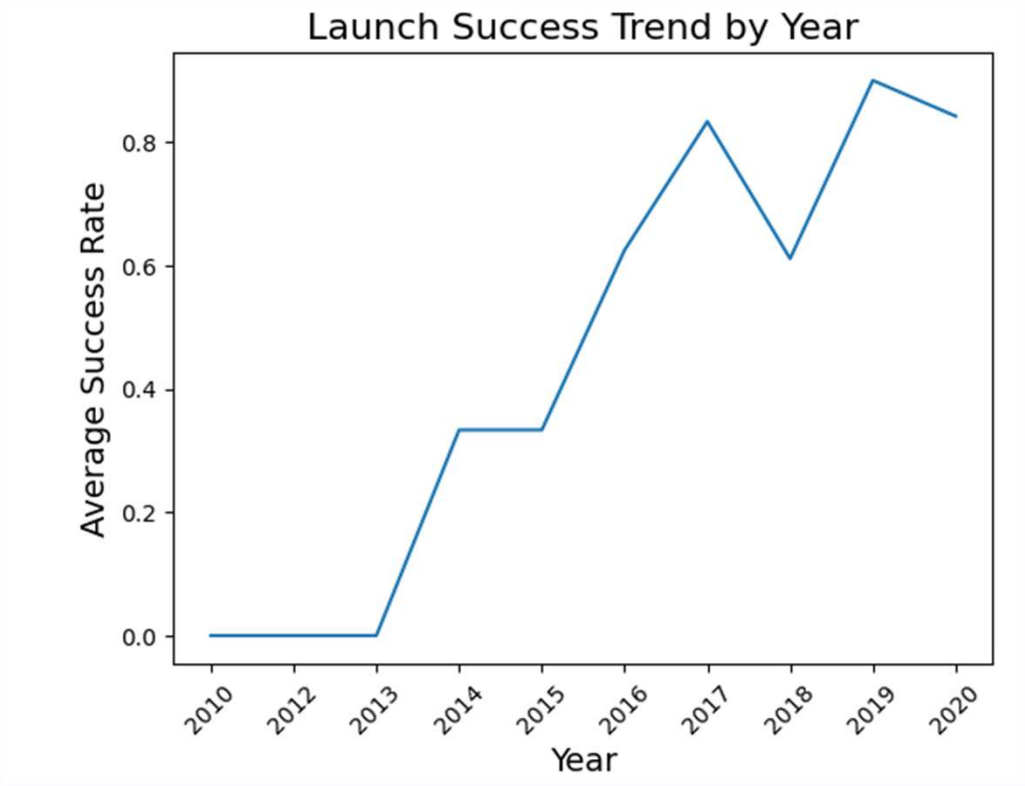
- This scatter plot compares Payload Mass across various Orbit Types, highlighting that Polar, LEO and ISS exhibit higher success rates. In contrast, GTO – a transitional orbit used for transfers - shows mixed outcomes, making direct comparisons more complex.



Launch Success Yearly Trend

Draft

- 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.



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

Draft

- 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

Draft

- 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

Draft

- 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

Draft

- 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

Draft

- 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

Section 3

Launch Sites Proximities Analysis

Launch sites with markers on a map

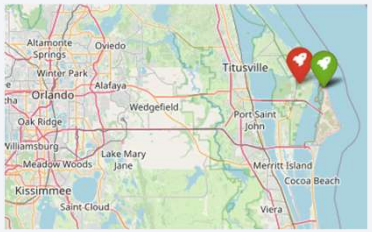
Draft

- Interactive Map of Launch Site Locations
  - This slide features a folium=power interactive map that plots all recorded launch site using geographic coordinates. Each site is marked with a clickable icon, allowing users to explore spatial distribution and site-specific metadata.
- Enhanced Context Trough Geolocation
  - By visualizing launch sites on a real-world map, stakeholders gain immediate geographic context – including proximity to coastlines, urban centers, and other infrastructure – which is critical for understanding logistical and environmental conditions.

Launch Outcomes on a map

Draft

- Site-Specific Outcome Visualization
  - This map highlights Cape Canaveral Air Force Station's SLC-40 launch site, overlaying individual launch outcomes directly onto the geographic location. Each mission is represented by a color-coded marker – green for success, red for failure – enabling rapid visual assessment of historical performance.
- Temporal and Operational Insights
  - By plotting outcomes over time, the map reveals patterns in launch reliability, including clusters of successful missions and periods of operational setbacks. This spatial temporal view supports root cause analysis and performance tracking
- Interactive Exploration of Mission Metadata
  - Users can click on each marker to access mission-specific details such as payload mass, booster version, launch date, and landing outcome. This interactivity transforms the map into a dynamic analytical tool for engineers, planners, and stakeholders.
- Foundation for Reusability and Recovery Planning
  - Mapping outcomes at the site level provides a baseline for evaluating reusability strategies, assessing environmental factors, and planning future recovery zones. It also supports comparative analysis with other launch sites.

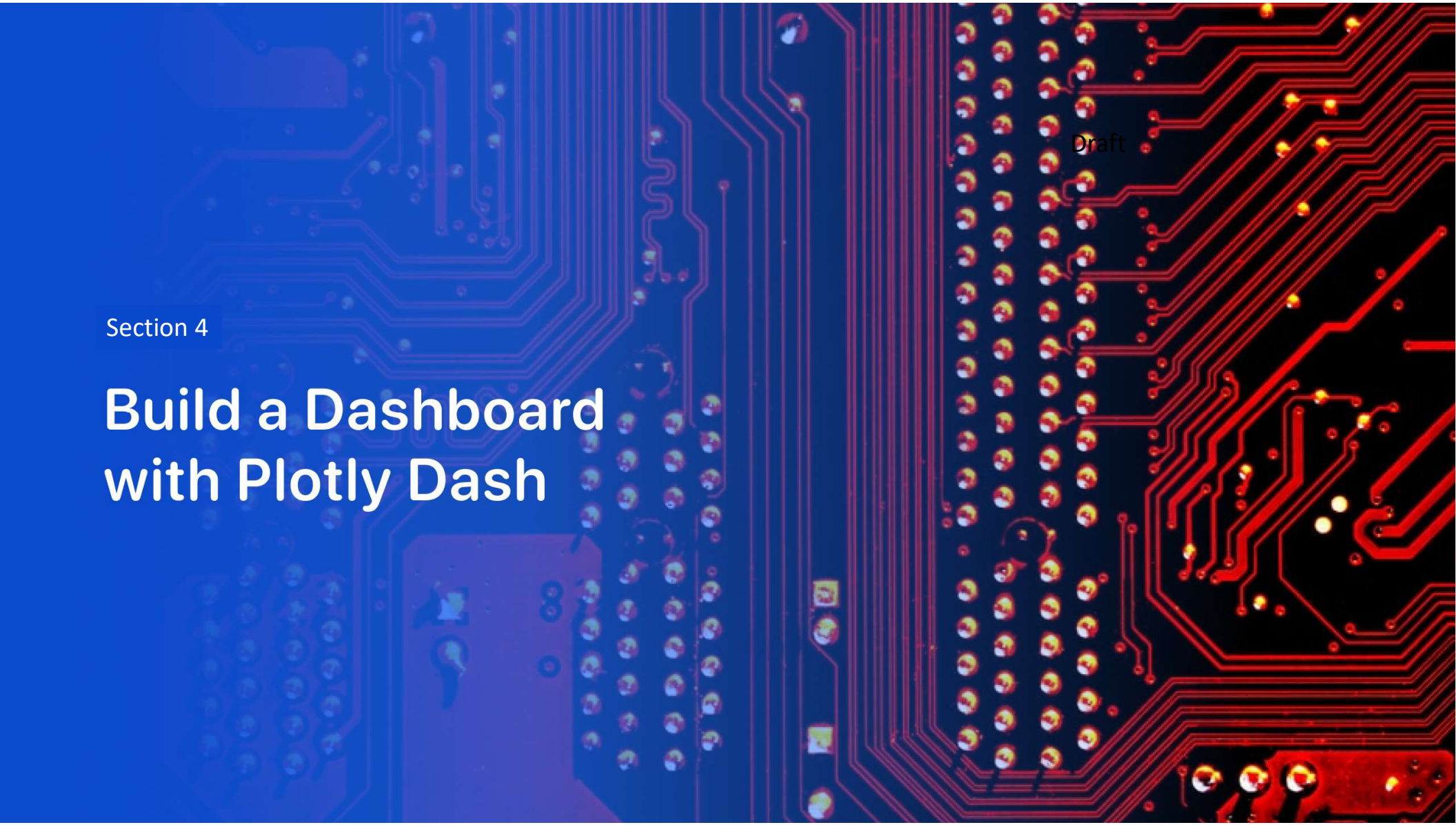


Distance between launch site and nearby proximities

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
```



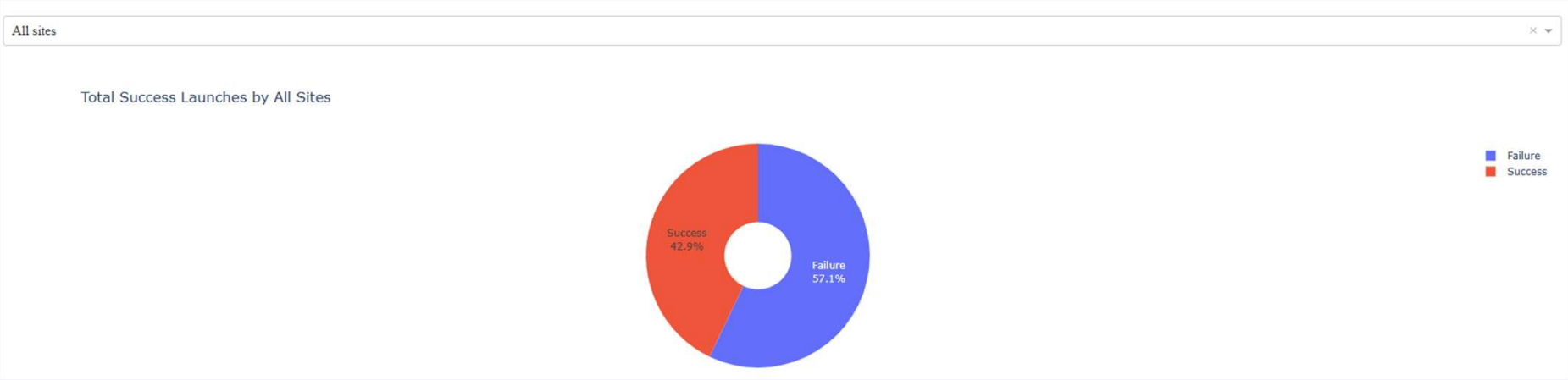
Section 4

Build a Dashboard with Plotly Dash

Total Success Launches by All site

Draft

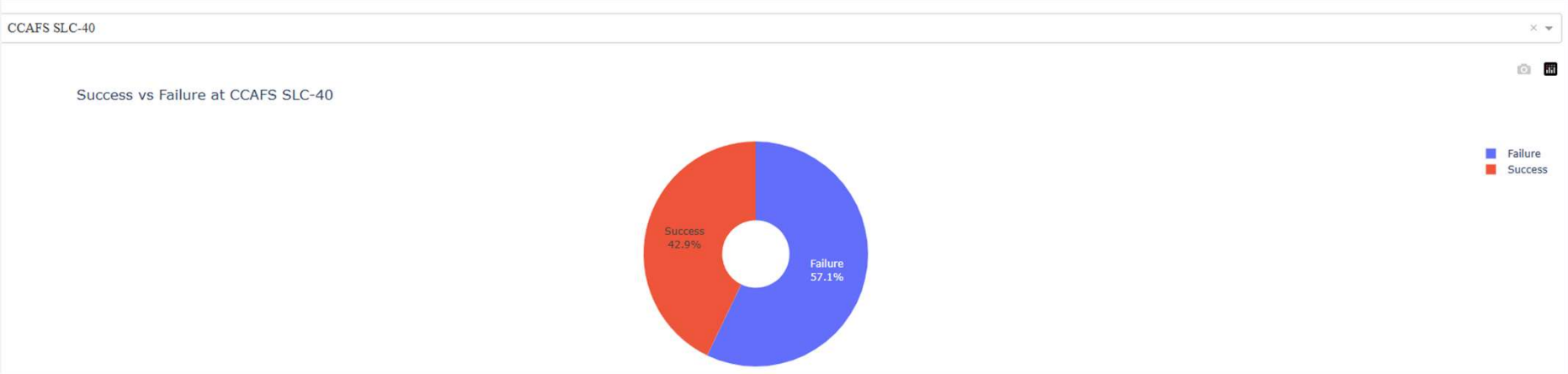
- This pie chart shows that at 42.9% of launches across all evaluated sites were successful, while 57.1% were classified as failures. However, it's important to note that many of these unsuccessful landings were intentional or plan, often part of test missions or low-priority recover attempts.



Launch Site with highest launch success ratio

Draft

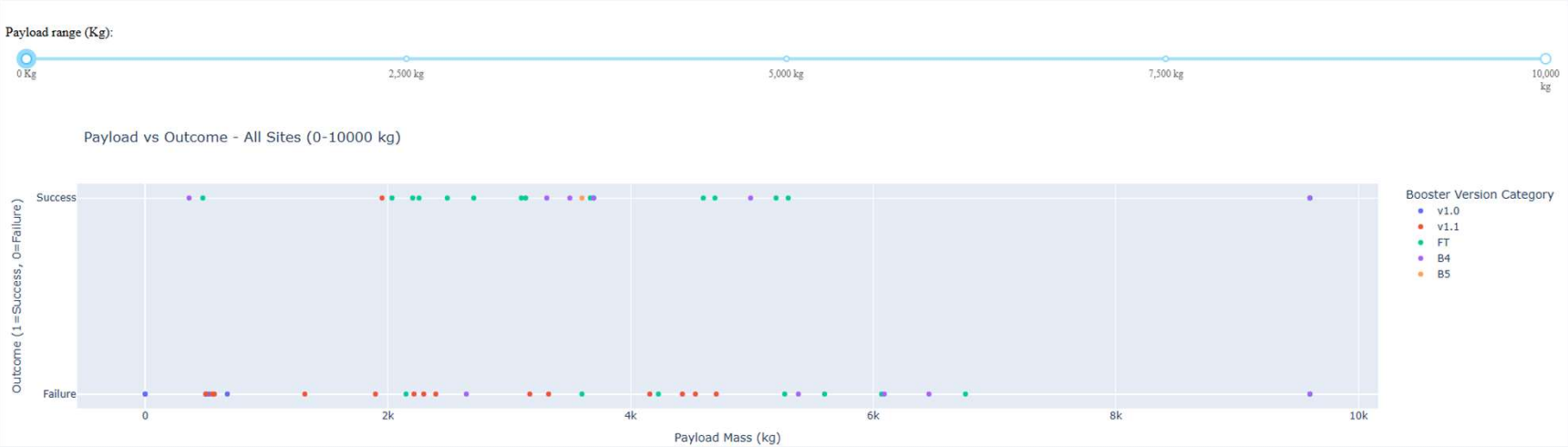
- This pie chart highlights CCAFS SLC-40 as the launch site with the highest success ratio, accounting for 42.9% of all successful launches among the evaluated locations.



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.



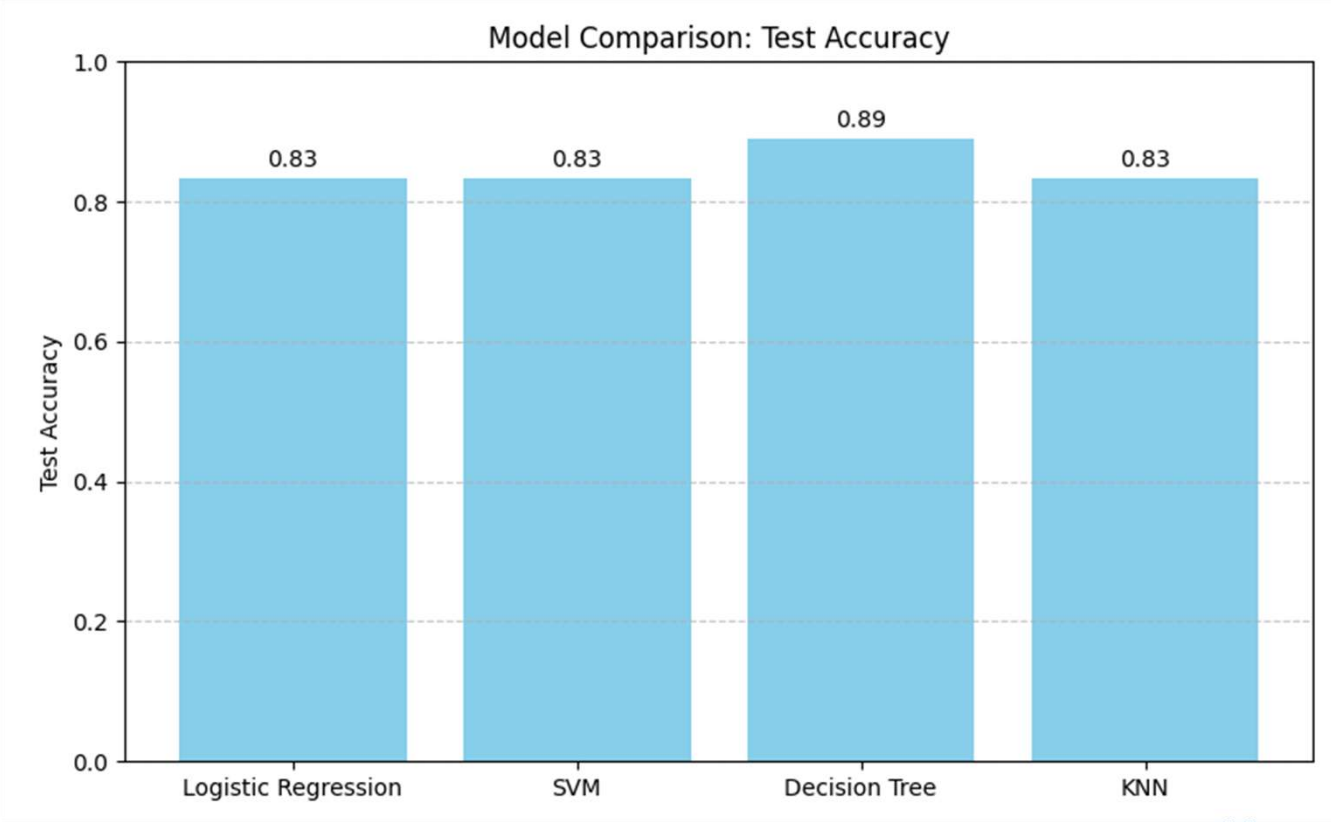
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Draft

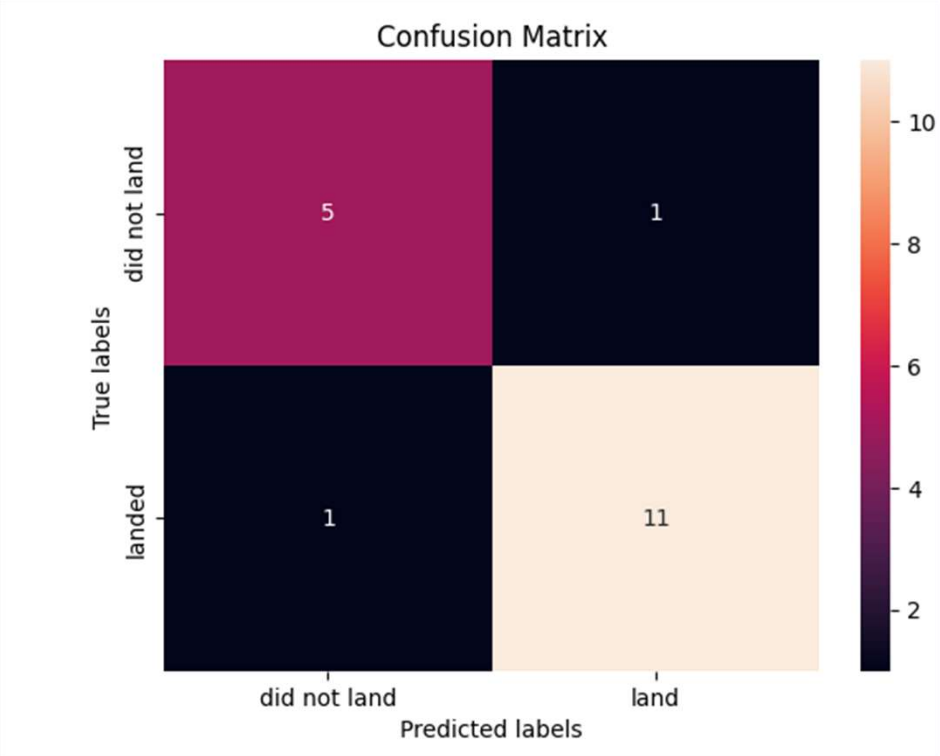
- Among all classification models evaluated, the Decision Tree achieved the highest accuracy, outperforming other approaches in predicting launch outcomes.



Confusion Matrix

Draft

- 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.



Conclusions

Draft

- Booster FT Reliability for Mid-Weight Payloads
  - Analysis revealed that the FT booster consistently achieved successful landings when paired with mid-range payload masses. This reliability profile suggests an optimal performance envelope, informing future booster-payload pairing strategies.
- CCAFS SLC-40 Top Performing Launch Site
  - Among all launch locations, Cape Canaveral Air Force station's SLC-40 stood out with the highest success rate. Its consistent performance underscores it's strategic value for mission planning and resource allocation.
- Decision Tree Model: Best-in-Class for Landing Prediction
  - After Evaluating multiple classifiers, the Decision Tree model delivered superior predictive accuracy for landing outcomes. Its interoperability and performance make it a strong candidate for operational forecasting.
- Confusion Matrix Validation
  - The model's confusion matrix confirmed high precision and recall, with minimal false positives or negatives. This reinforces confidence in the model's reliability for real-world development and stakeholder reporting.
- Reusability Milestone: Late 2015 Onward
  - Historical data pinpointed late 2015 as the inflection point for successful booster landings, marking the beginning of a new era in launch vehicle reusability and cost-efficiency.
- NASA missions: Major Contributor to Payload Volume
  - NASA-sponsored missions accounted for substantial share of total payload mass, highlighting the agency's pivotal role in driving launch activity and scientific advancement.
- Strategic Insights for Future Launch Planning
  - These data-driven findings support informed decision-making around launch site selection, booster configuration, payload optimization, and recovery planning – laying the groundwork for scalable, cost-effective mission design.

Strategic Insights

Draft

- Reusability Trends Align with Payload Efficiency
  - The emergence of successful booster landings post-2015 coincides with payload mass optimization and mission planning maturity. This suggests a growing synergy between vehicle design, payload engineering, and recovery strategy.
- Booster Performance Varies by Payload Class
  - Booster FT's reliability with mid-weight payloads highlights the importance of matching booster configurations to mission profiles. Future planning should consider payload class as a key variable in booster selection and risk modeling.
- Launch Site Selection Impacts Mission Success
  - CCAFS SLC-40's standout performance reinforces the need for site-specific reliability metrics in launch planning. Environmental factors, infrastructure readiness, and historical success rates should inform site prioritization.
- Machine Learning Enhances Predictive Planning
  - The Decision Tree model's strong performance demonstrates the value of interpretable ML in operational forecasting. Integrating predictive analytics into pre-launch assessments can reduce risk and improve resource allocation.

Strategic Insights Cont.

Draft

- NASA's Payload Dominance Signals Strategic Partnerships
  - NASA's significant payload contribution underscores the importance of institutional partnerships in sustaining launch cadence and scientific output. Future collaborations should leverage this momentum for joint innovation.

- Spatial Mapping Enables Recovery Optimization
  - Folium-based geospatial visualizations offer a powerful lens for planning recovery zones, assessing launch site logistics, and integrating external data layers (e.g., weather, terrain, transport corridors).

- Dashboard Interactivity Drives Stakeholder Engagement
  - The use of dropdowns, sliders, and dynamic charts transforms static data into exploratory tools. This interactivity fosters deeper stakeholder engagement and supports data-driven decision-making across technical and executive audiences.

Closing Summary & Strategic Outlook

Draft

- Interactive Dashboards Drive Stakeholder Engagement
  - The integration of dropdowns, sliders, and dynamic charts transforms strategic data into actionable insights, empowering users to explore launch performance, payload dynamics, and site-specific trends.

- Geospatial Mapping Enhances Operational Planning
  - Folium-based visualizations provide spatial context for launch sites and outcomes, supporting recover logistics, infrastructure analysis, and future site selection.

- Machine Learning Enables Predictive foresight
  - The Decision Tree model's strong performance in landing outcome prediction demonstrates the value of interpretable ML for mission planning, risk mitigation, and resource optimization.
- Historical Trends Inform Reusability Strategy
  - The post-2015 shift toward successful booster landings marks a pivotal evolution in launch economics and sustainability, guiding future investment in reusable technologies.
- NASA's Payload Dominance highlights Strategic Partnerships
  - NASA's significant payload contributions reinforce the importance of institutional collaboration in driving launch cadence and scientific impact.
- Data-Driven Insights Support Scalable Innovation
  - From payload to booster matching to launch site reliability, the dashboard delivers actionable intelligence for refinding launch configurations, enhancing recover methods, and planning future missions.

Appendix

- All code can be found in the GitHub repository for this project and there are links to it throughout this file.
  - PMDataGeek25/Capstone-Project: This is my repository for my Capstone Project

Thank you!

Draft