IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Quentin Stepp
04.26.2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Data Collection and Preparation

- Collected via SpaceX's API and web scraping from Wikipedia using Python's requests and BeautifulSoup.
- Data cleaned, missing values addressed, and a new landing outcome column added.
- Public, reproducible datasets; sources mitigate risk of improper collection.

Exploratory Data Analysis (EDA)

- Visualizations explored relationships among flight number, payload mass, launch site, orbit type, and success rate.
- SQL queries provided additional insights, such as average payload by booster version and landing outcome counts.

Interactive Visual Analytics

- **Folium** maps visualized launch site locations, successes, and proximity to coastlines.
- **Plotly Dash** enabled dynamic, user-driven data exploration.

Predictive Analysis

- Built four classification models: Logistic Regression, SVM, Decision Tree, and KNN.
- Applied Grid Search for hyperparameter optimization.
- Achieved ~83% prediction accuracy for landing outcomes.

Key Insights

- Launch site, payload mass, and booster type significantly impact landing success.
- Interactive tools allow easy identification of high-performing launch sites, like KSC LC-39A.

Introduction

Project Scenario

- The commercial space race is booming, with companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX leading innovation.
- SpaceX has lowered launch costs dramatically by reusing the first stage of its Falcon 9 rocket.

Problem Statement

- Successful recovery of the Falcon 9's first stage is crucial to lowering launch costs (from ~$165M to ~$62M per launch).
- Predicting first-stage landing success can directly impact cost estimation and operational decisions.

Our Role

- As a data scientist at SpaceY, a competitor to SpaceX, you must predict whether the first stage will land successfully.
- You'll leverage public SpaceX data, create interactive dashboards, and train machine learning models to predict first-stage recovery.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data is collected both by requests to SpaceX's API and web scraping of a Wikipedia which details SpaceX launch data.
- Perform data wrangling
  - The data was cleaned of any missing values, and a column was added with a dummy variable indicating whether a landing was successful or not.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four classification models were used (Logistic Regression, an SVM, a Decision Tree, and KNN). Their relative performances were analyzed.

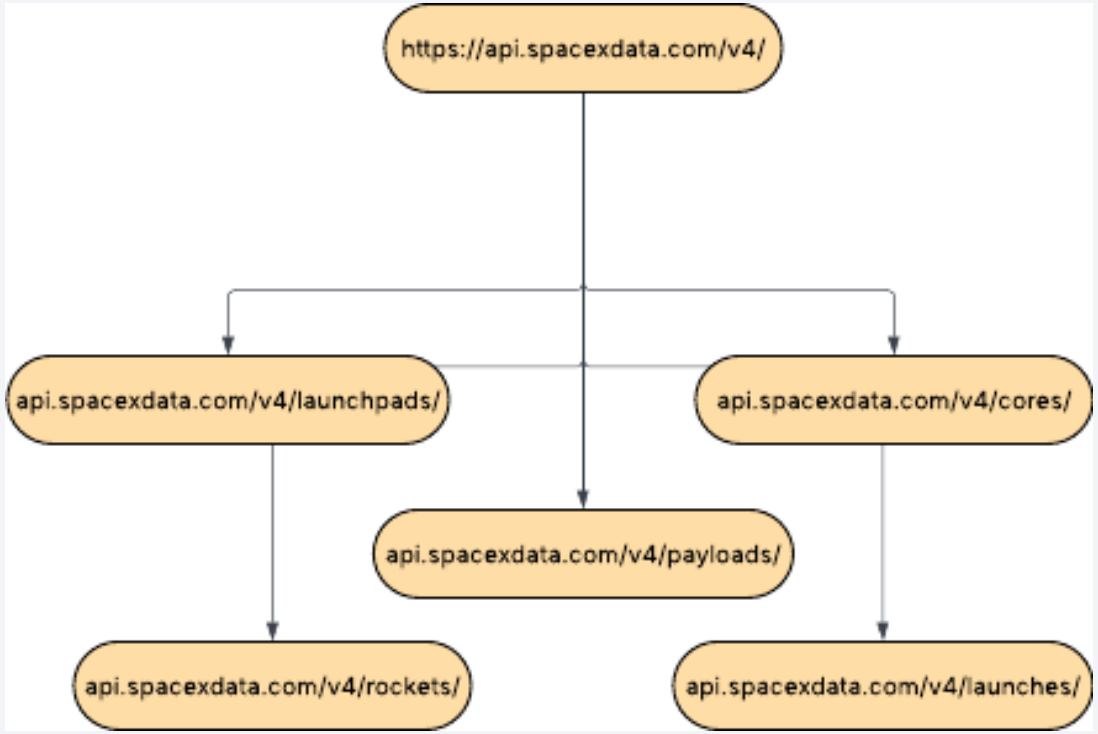
Data Collection

- Data were collected by both web scraping and requests to SpaceX's data API.
- All data used for this project were publically available and all submitted work is reproducible.
- The use of both publicly sourced data from Wikipedia and data from SpaceX themselves mitigates risk of improper or dishonest collection.
- All data imports are available in the attached GitHub repository links.

Data Collection – SpaceX API

- Data is collected using requests directly to SpaceX's API. The relevant data is assembled into a single, usable data set.

- GitHub URL



Data Collection - Scraping

- Web Scraping with BeautifulSoup: Used Python's requests to fetch the Wikipedia page on Falcon 9 and Falcon Heavy launches, then parsed the HTML with BeautifulSoup to extract launch data from tables.
- From here, minimal data cleaning and wrangling was required to transform the data into a usable form.
- GitHub URL

Data Wrangling

- The data wrangling process for this project consisted of some minor data manipulation and augmentation to facilitate further analysis and predictive analysis.
- The data set was analyzed for NaN values, which were ultimately left in.
- A landing outcome column was added, providing a binary variable indicating whether a laund landed successfully.
- On average 2/3 of all launches landed successfully.
- GitHub URL

EDA with Data Visualization

- The EDA process produced visualizations that provided key data insights.
- These include comparisons of:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Orbit Type vs. Launch Success Rate
  - Flight Number vs. Orbit Type
  - Payload Mass vs. Orbit Type
  - Success Rate vs. Time
- The visualizations allowed for better feature engineering and data selection to be used in the predictive analysis.
- GitHub URL

EDA with SQL

- The ensuing dataset was further analyzed with SQL.
- Important statistics were displayed, such as total unique launch sites, average payload mass by booster version, booster versions used within a payload mass range, or a table of landing outcome counts.
- These insights are uniquely accessible with SQL, which allowed for quick analysis of key features.
- GitHub URL

Build an Interactive Map with Folium

- I used a Folium interactive map to visualize the location of each launch site and compare the relative success of each.
- I added location markers and radii for each launch site to bring their physical locations into my analysis.
- I then visualized each launch from these sites and whether they were successes or failures.
- Finally, I calculated the distances between the launch sites and their nearest coastlines in order to demonstrate their proximity to the water.
- GitHub URL

Build a Dashboard with Plotly Dash

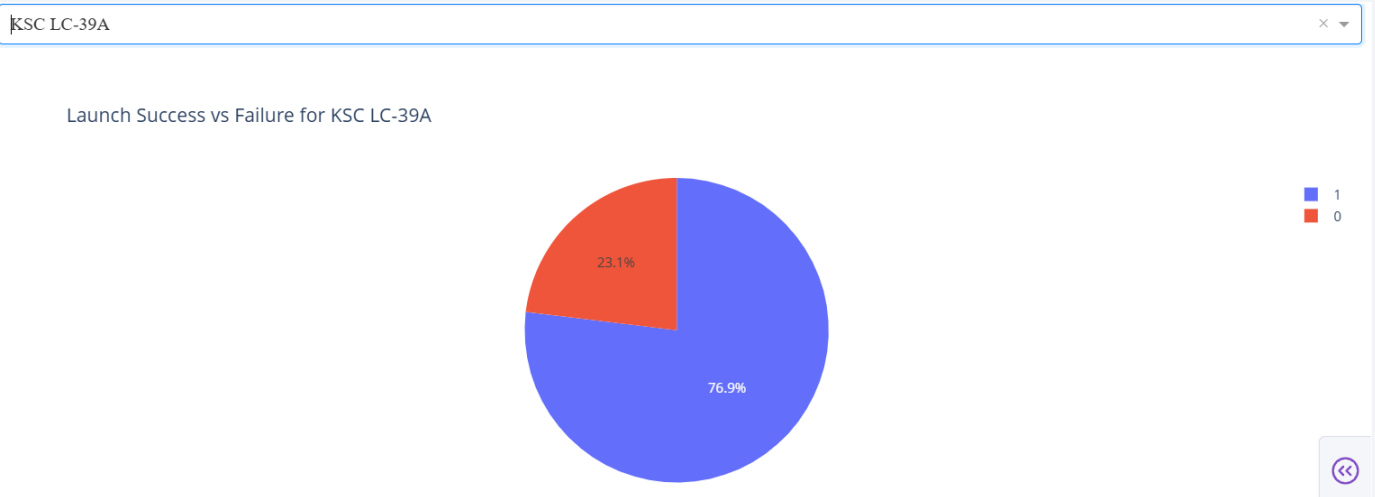
- In my Plotly Dashboard, I created a tool to visualize the success rate of launch sites.
- This allows users to manually compare the success / failure rate of each individual site or the aggregate of all 4 sites.
- The dashboard also includes a scatterplot showing the Payload Mass of each launch, relative to their success or failure.
    - I included this to provide a tool to compare the success of launches based on their payload mass.
- GitHub URL

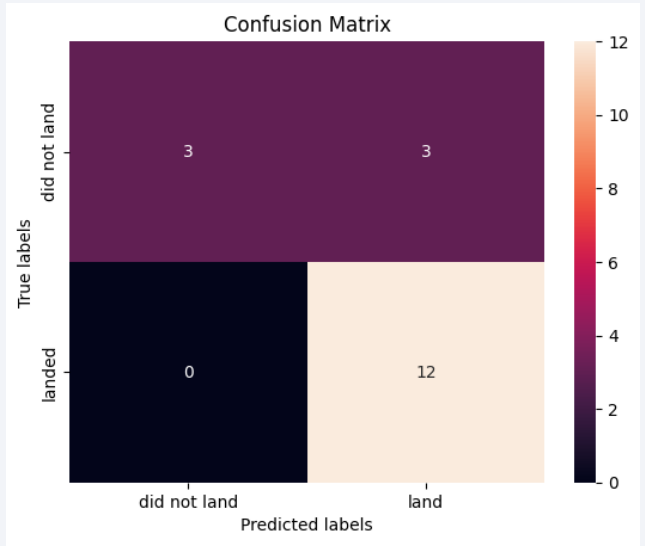
Predictive Analysis (Classification)

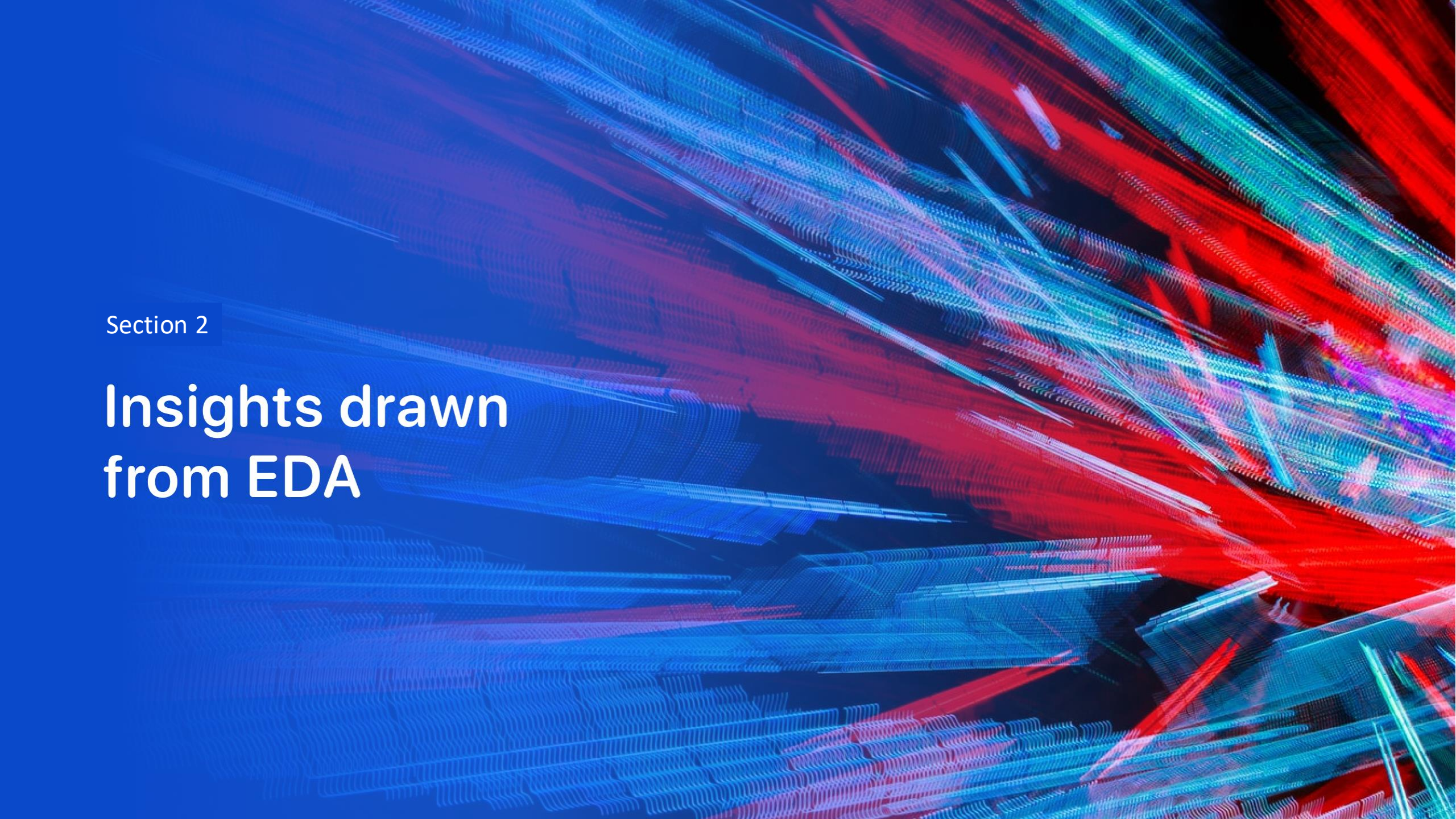
- Four classification models were used to perform predictive analysis:
  - Logarithmic Regression
  - SVM
  - Decision Tree
  - KNN
- The data was first scaled and split into training and test sets (20% validation size).
- A Grid Search was conducted with each model to find the best perfroming hyper-parameters.
- Model performance was assessed using an accuracy score on the test data.
- GitHub URL

Results

- The EDA process revealed connections between the mass of the rocket, the launch site, and the type of booster with the chances of success. This allowed better feature selection for further analysis.
- Interactive analytics allows a user to explore the data themselves. For example, one could identify that the KSC LC-39A launch site has the highest success rate without any coding.
- The ensuing models were able to predict whether a launch would land successfully with ~83% accuracy, which is a valuable insight for SpaceY.

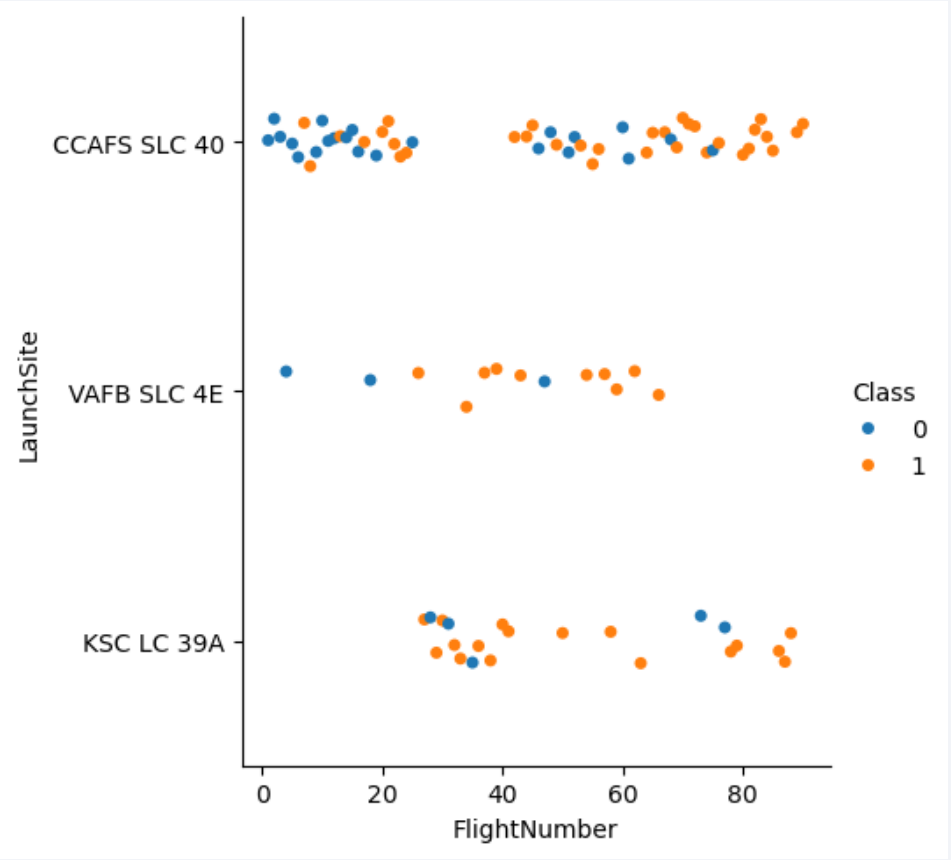





Section 2
Insights drawn from EDA

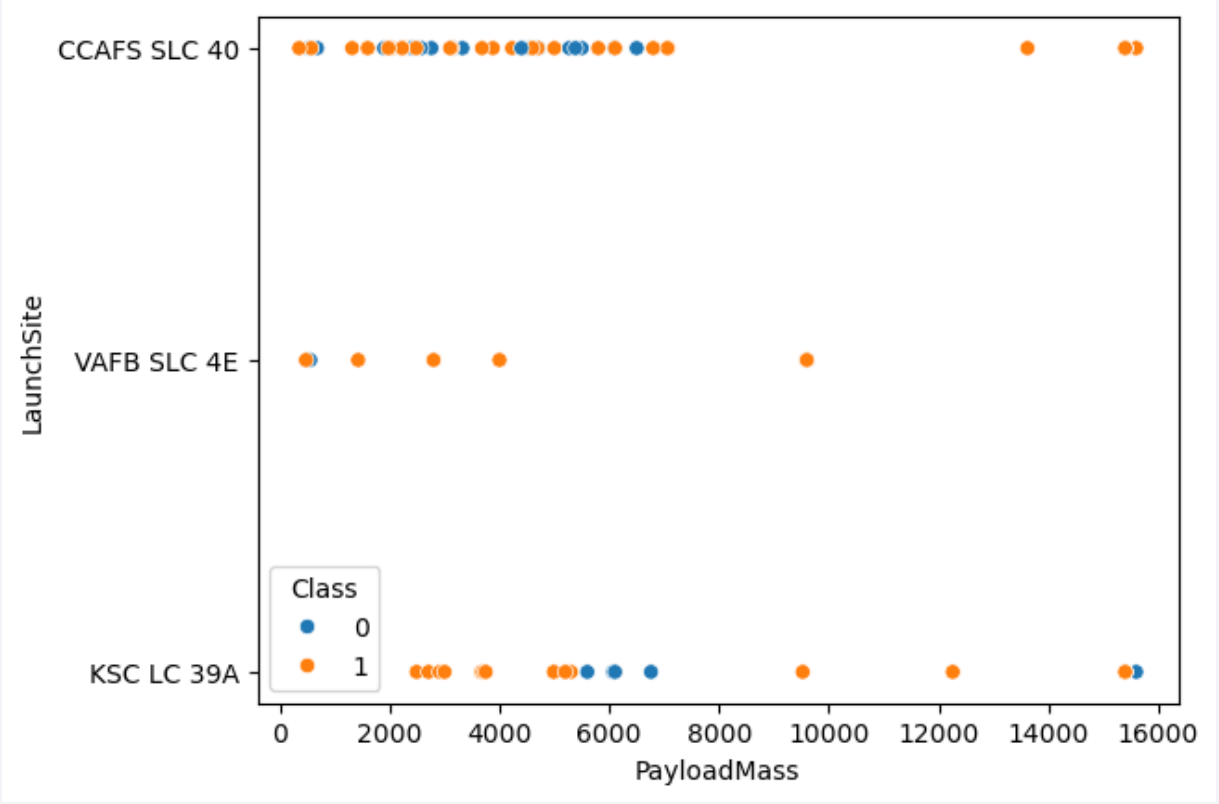
Flight Number vs. Launch Site

- This plot shows ordinal flight numbers plotted against their relative launch sites.
- Each dot is colored based on their being a successful or failed launch, orange being successful.
- It appears that the VAFB SLC 4E site has the highest proportion of successful launches.
- There seem to be a higher success rate in later flight numbers, showing overall SpaceX improvement.



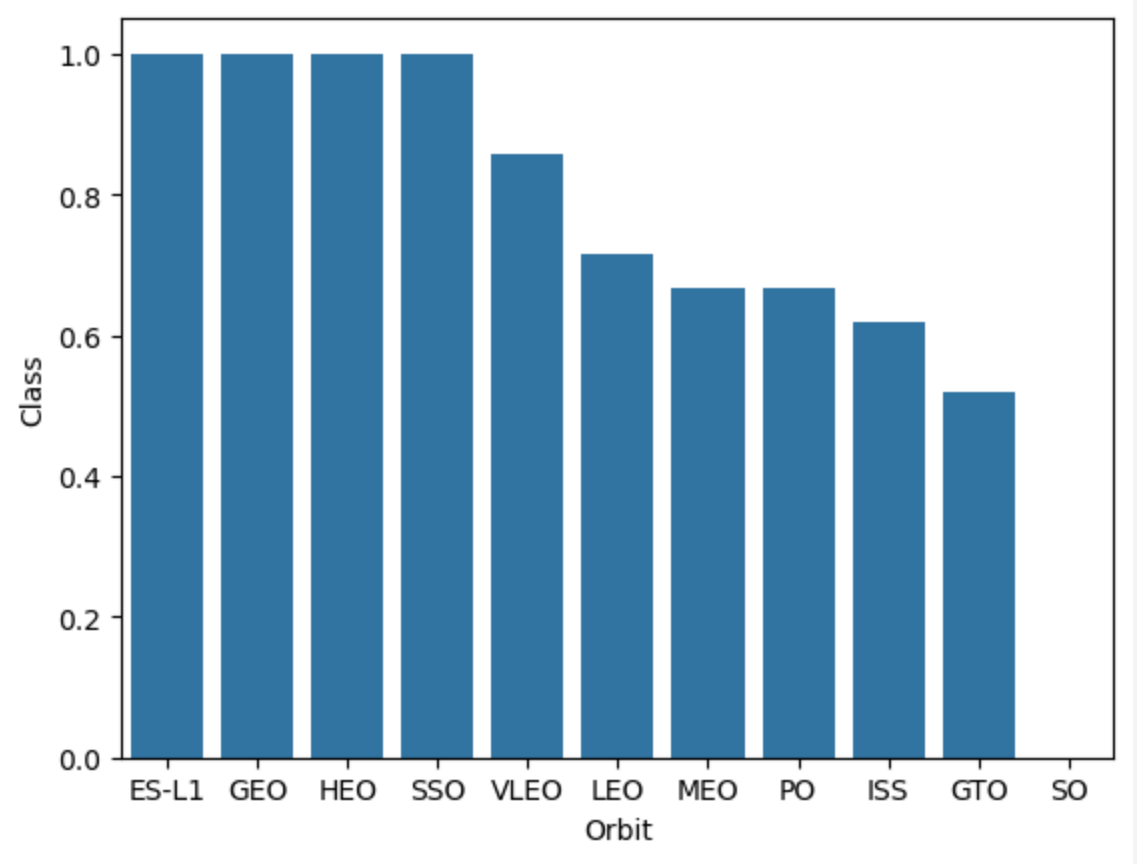
Payload vs. Launch Site

- This graph shows each launch's Payload Mass plotted against their Launch Site.
- Each dot is colored based on their being a successful or failed launch, orange being successful.
- It seems like there may be a slightly better chance of success with a heavier payload but it is difficult to see this conclusively here.



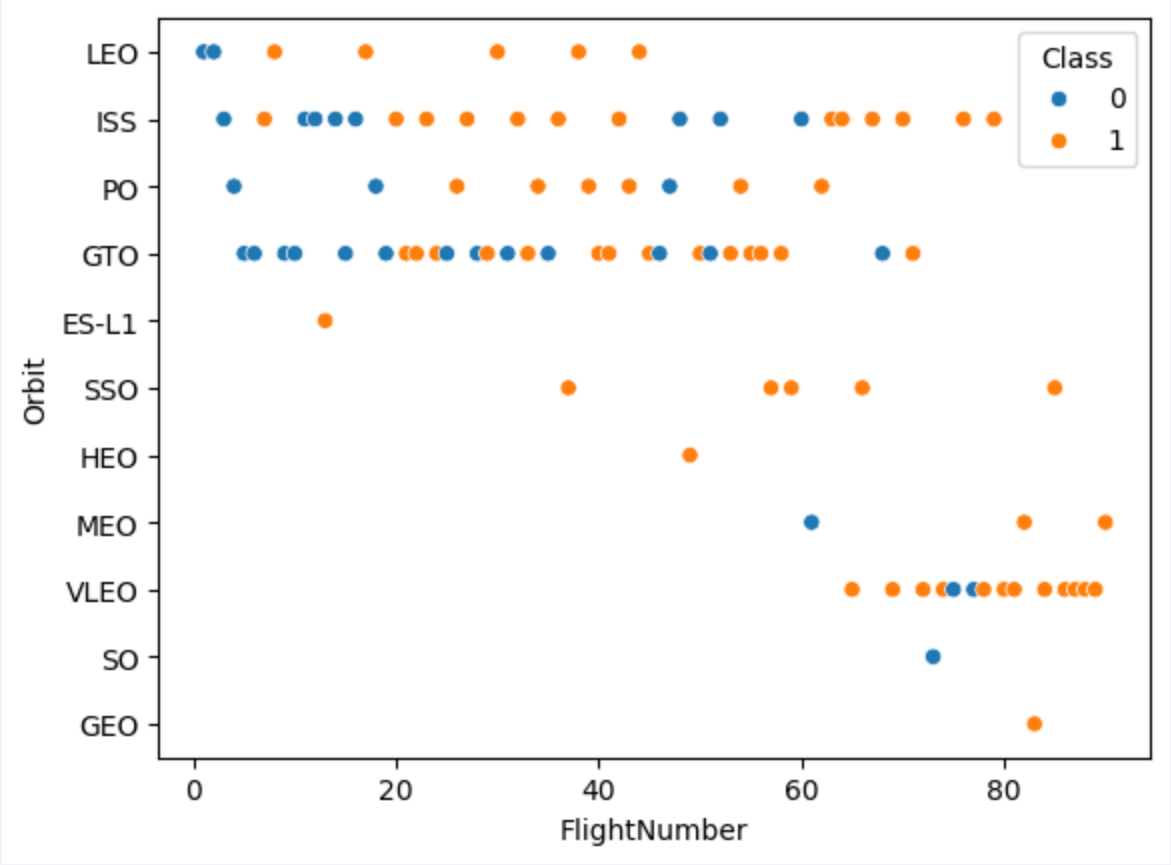
Success Rate vs. Orbit Type

- This chart plots the relative success rate of each Orbit type.
- ES-L1, GEO, HEO, and SSO orbit types have perfect success rates.
- GTO Orbits appear to have the lowest success rates



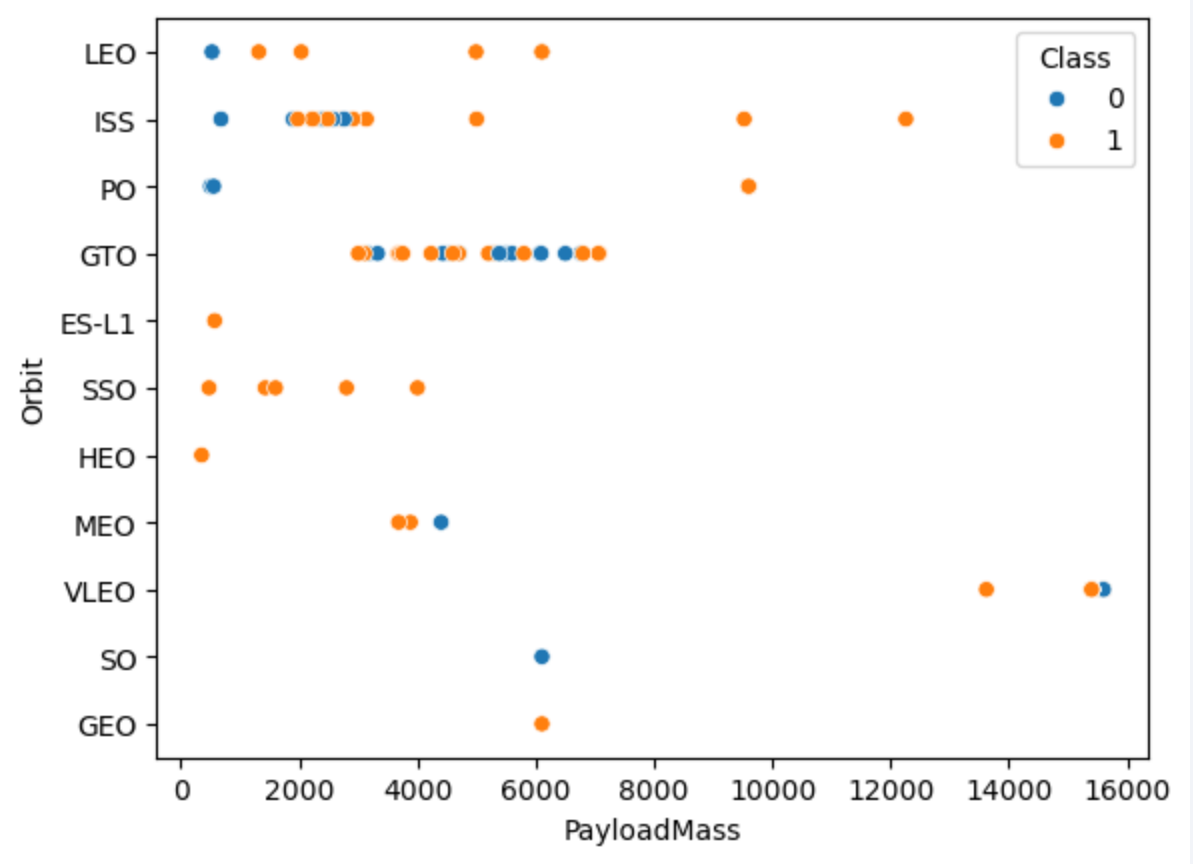
Flight Number vs. Orbit Type

- This graph shows each ordinal Flight Number plotted against Orbit types.
- Each dot is colored based on their being a successful or failed launch, orange being successful.
- These data show that the Orbit types with perfect success rates actually had relatively few data points.



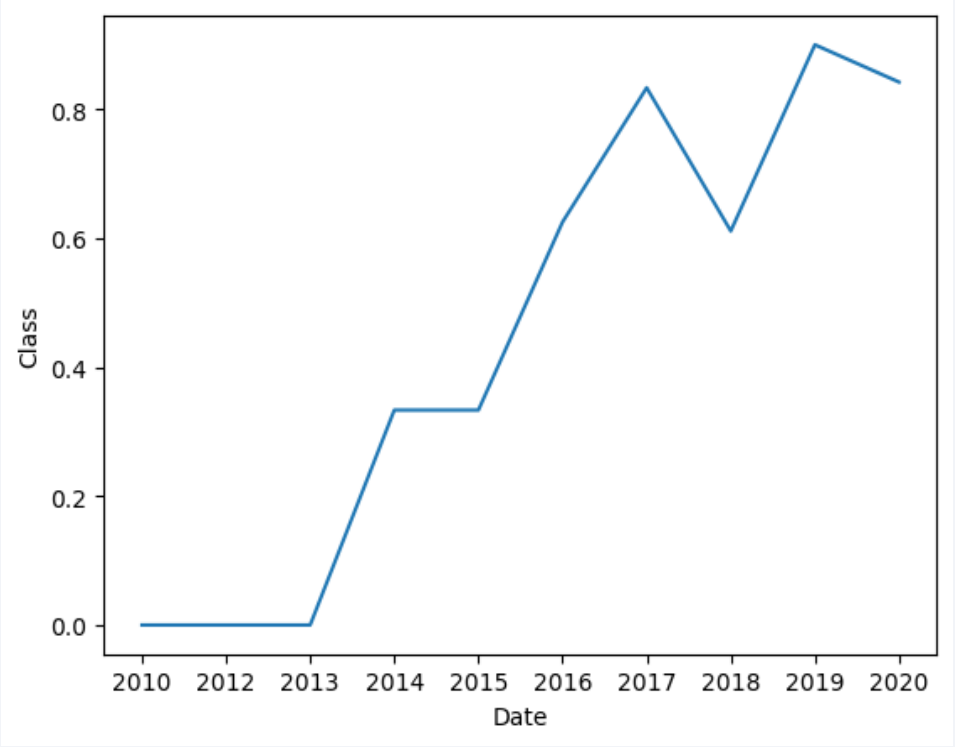
Payload vs. Orbit Type

- This plot shows Payload Mass relative to Orbit type, with points shaded to indicate success or failure.
- VLEO launches appear to be the heaviest while SSO launches are on the lighter side.



Launch Success Yearly Trend

- This chart shows Launch Success Rate over time.
- There appears to be a steady increase in launch success rates.



All Launch Site Names

- Here, the names of all unique launch sites were extracted from a database using the query:
  - SELECT DISTINCT Launch_Site FROM SPACEXTABLE
- The resulting table is attached here. I selected all distinct launch site names from the relevant table.

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

Launch Site Names Begin with 'CCA'

- I found 5 records where launch sites begin with `CCA` using the query:
  - SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
- Here, I selected all rows from the relevant table, conditioned on the Launch Site value beginning with 'CCA'. I then limited the results to 5 entries.

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

- The total payload carried by boosters from NASA was calculated with the query:
  - SELECT SUM(PAYLOAD_MASS__KG_) AS Payload_Total FROM SPACEXTABLE WHERE Customer LIKE 'NASA%'
- Here, I select the sum of the Payload Mass attribute from the relevant table on the condition that the Customer column includes 'NASA'
- The total amount was 99,980 kg.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated with the query:
  - SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
- Here, I selected the average Payload Mass value from the relevant table, conditoned on the Booster Version being 'F9 v1.1'
- The average Payload Mass for this booster was 2,534.66

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad were found using the query:
  - SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
- This query selects the minimum Date from the relevant table, conditioned on the Landing Outcome being successful and on a ground pad.
- The resulting date was 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 was listed using the query:
  - SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 AND Landing_Outcome = 'Success (drone ship)'

- This query selects a Booster Version from the relevant table, conditioned on the Payload Mass being between 4,000 and 6,000 kg and the landing outcome being successful and on a drone ship.

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes was calculated using the query:
  - SELECT Mission_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE GROUP BY Mission_Outcome
- This query selects the Mission Outcone name and the count of all outcomes grouped by their Mission Outcome.

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass were listed with the query:
  - SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
- This query selects a Booster Version from the relevant table conditioned on only those boosters that carried the maximum Payload Mass. This was accomplished using a subquery.

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

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed below.
- This chart shows that there were just two launches that failed in this manner in 2015, both of which were at the same launch site and only 4 months apart.

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

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, are ranked in descending order below.
- A plurality of launches did not attempt a landing. Most successful landings were on a drone ship.

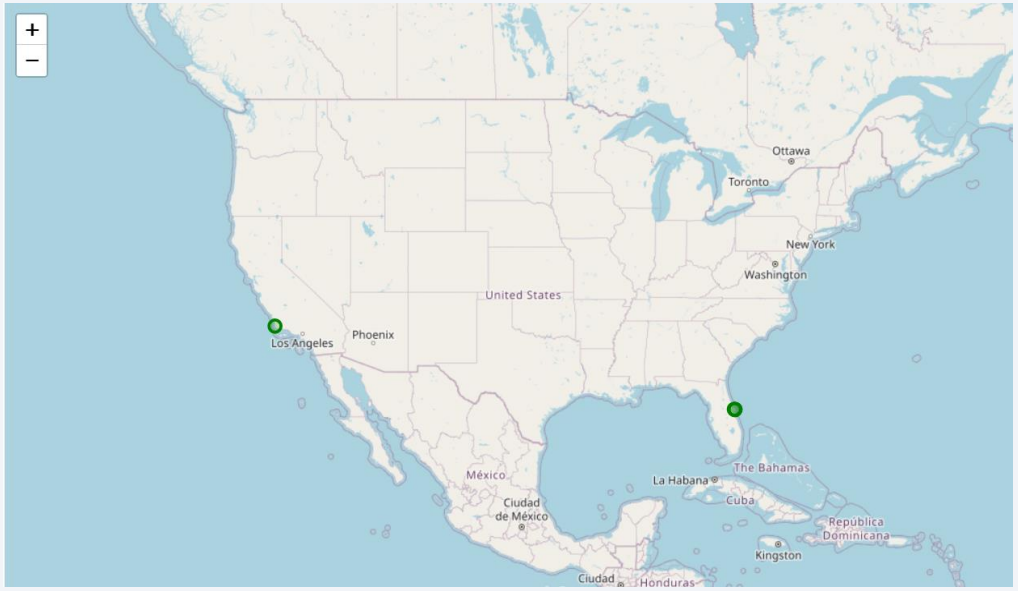
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

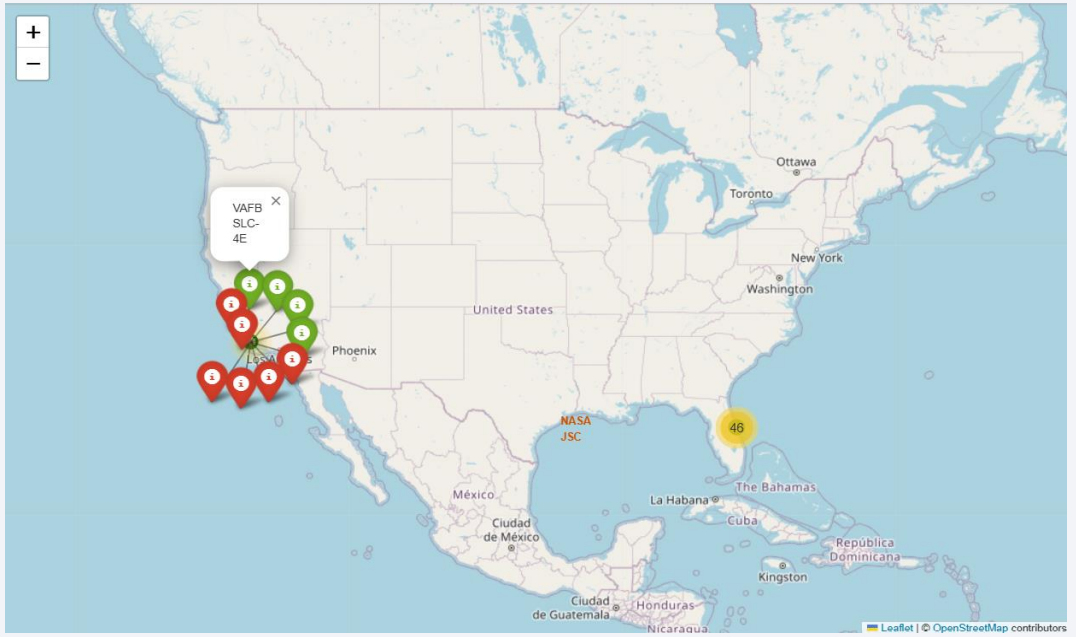
All SpaceX Launch Sites, Mapped

- SpaceX's launch pads are all coastally located, with three in Florida and one in California.
- Given the tendency for SpaceX to land their crafts in the ocean, it follows that their launch sites would be located coastally. This aligns with other historic launch sites, such as Cape Canaveral in Florida and the Johnson Space Center in Houston.



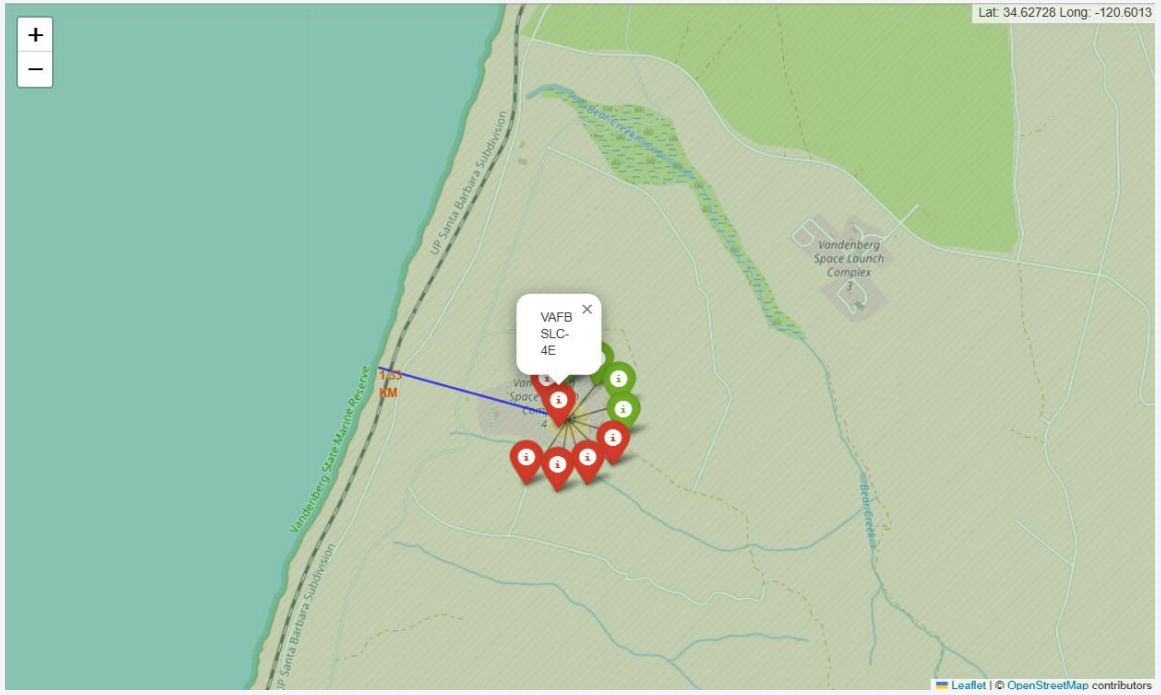
Launch Outcomes by Site

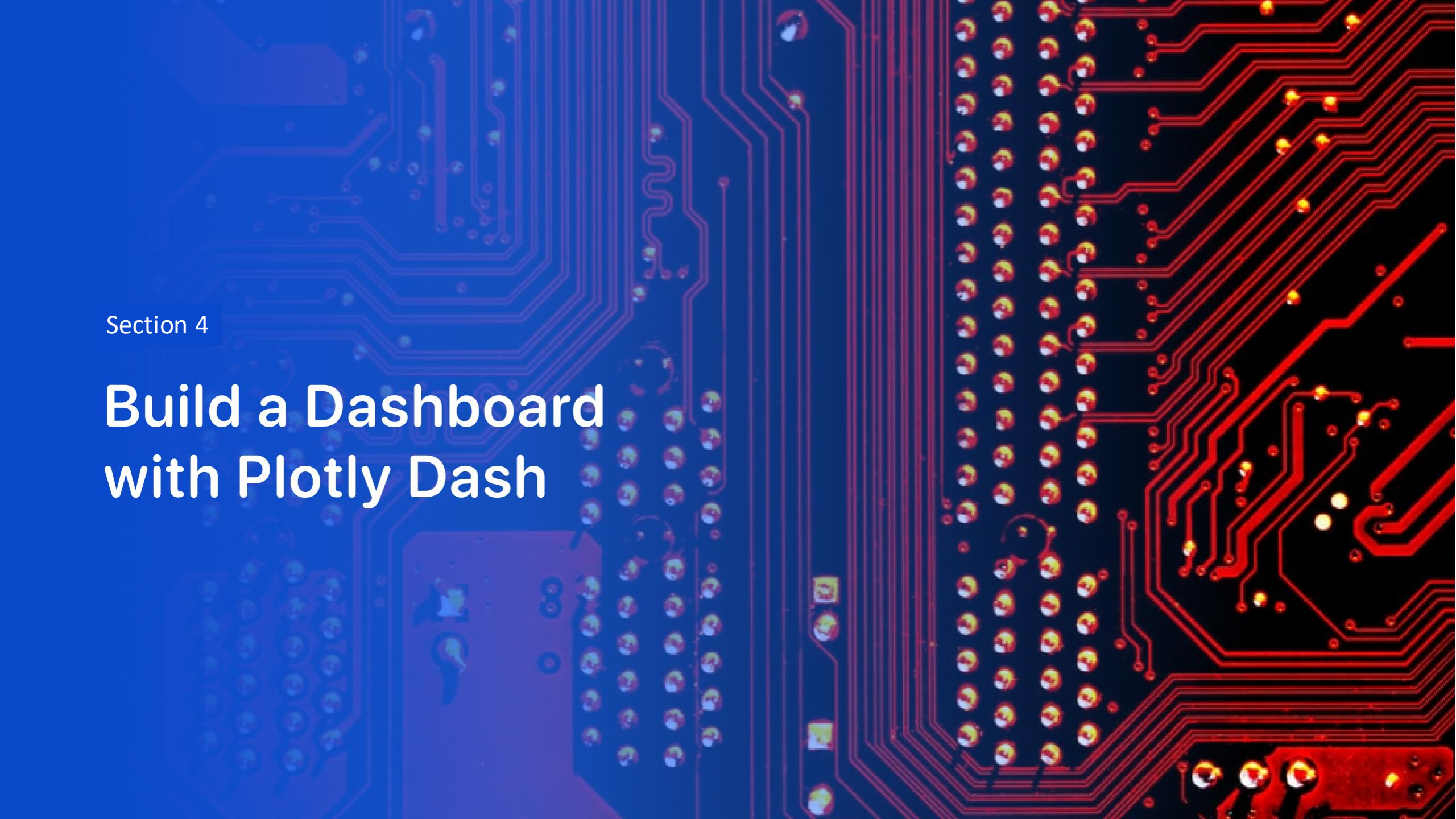
- This map displays each launch at their respective site with color designation for successful vs. unsuccessful launches.
- The KSC LC-39A Launch Site appears to be more conducive to successful launches, but no conclusive findind can be drawn from this visualization alone.



Launch Site Distance to Coastline

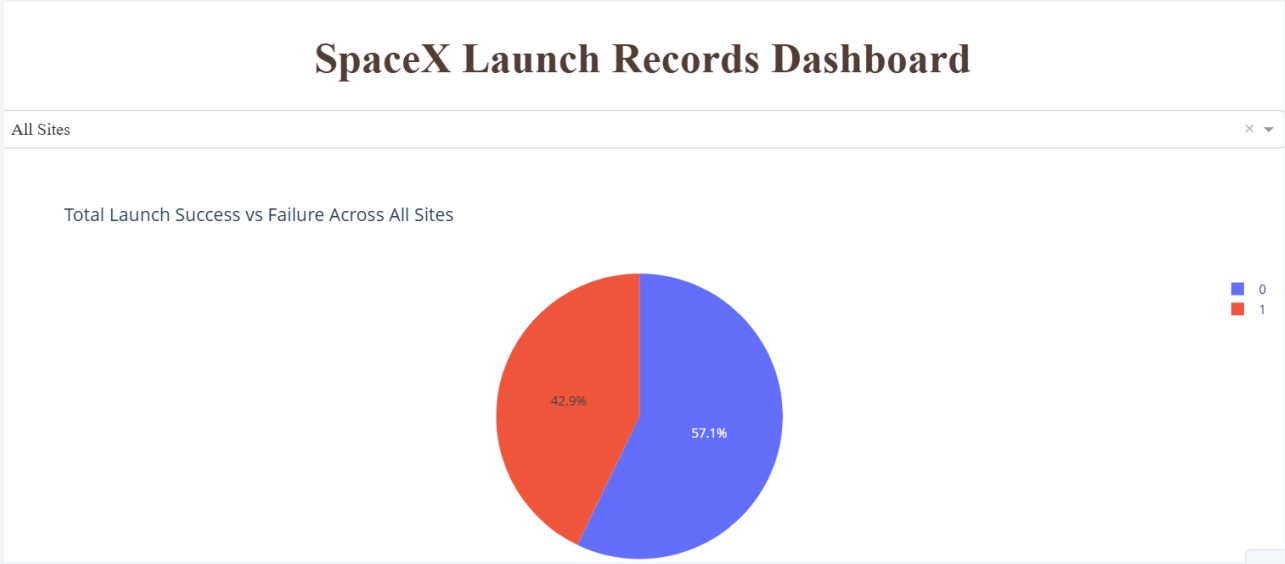
- This visualization displats the distance from the VAFB SLC-4E Launch Site to its nearest coastline.
- As previouly stated, SpaceX launch sites are primarily located along coasts. This site is a mere1.33 km away from the water.




Section 4
Build a Dashboard with Plotly Dash

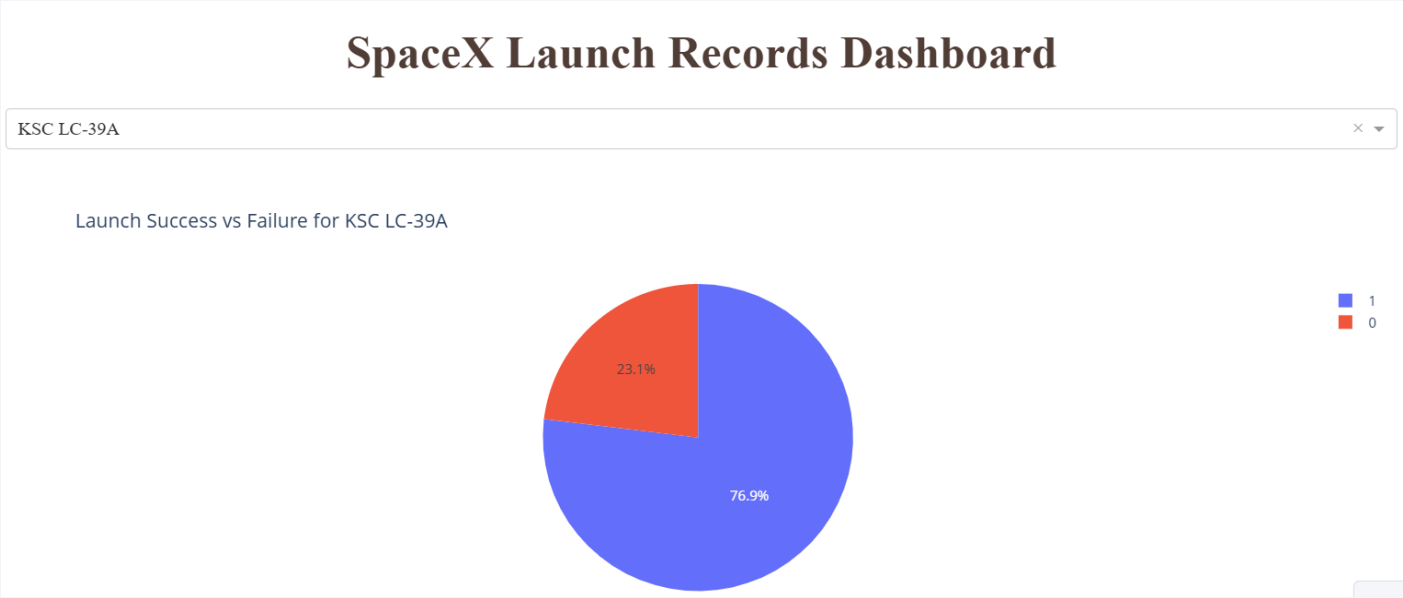
Launch Success Rate Dashboard

- The launch success rate across all sites was 57.1%, shown below in an interactive dashboard.
- This provides a useful overview of SpaceX performance since its inception.



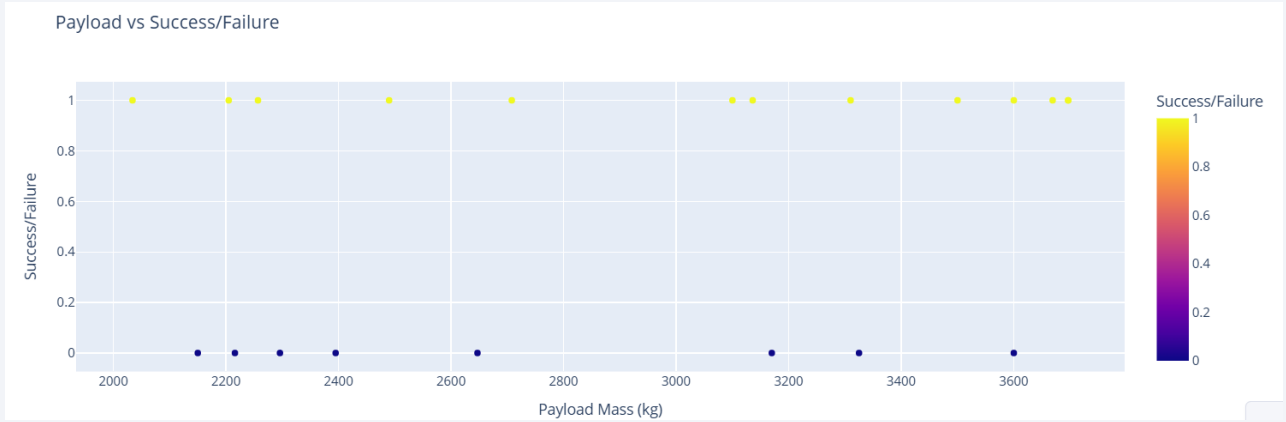
Best Launch Site

- Here, we see that KSC LC-39A is the launch site with the highest rate of success. Its success rate is significantly higher than the mean.
- This insight would not have been easily accessible without a dashboard but can be displayed here with a single click.



Payload vs Success/Failure

- Below is an interactive graph that shows whether launches were successful compared to their Payload Mass.
- This tool allows a user to select a range of Payload Mass values to display.
- The example below is from the range 2,000 to 4,000 kg, showing the success rate of smaller payloads.

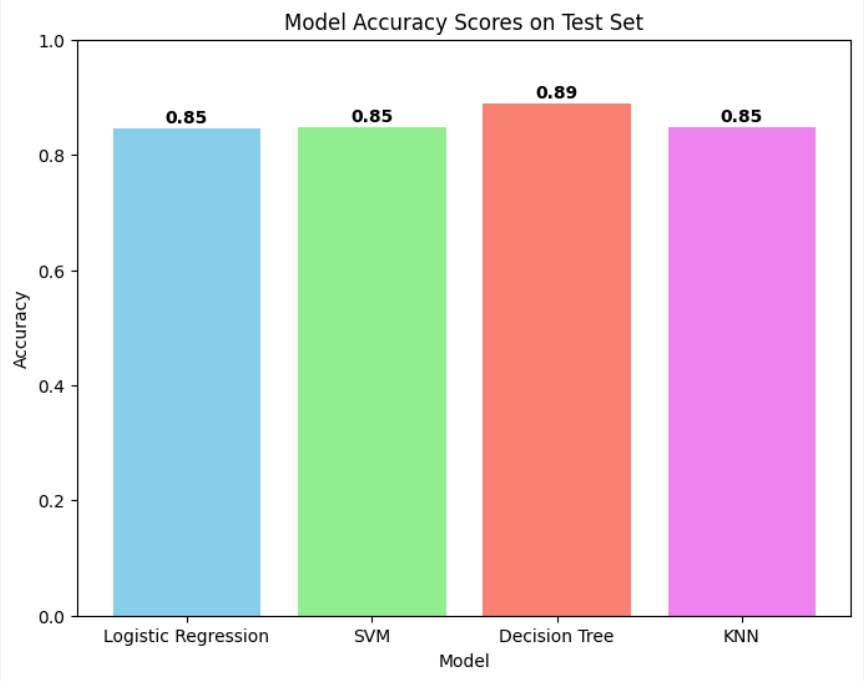


Section 5

Predictive Analysis (Classification)

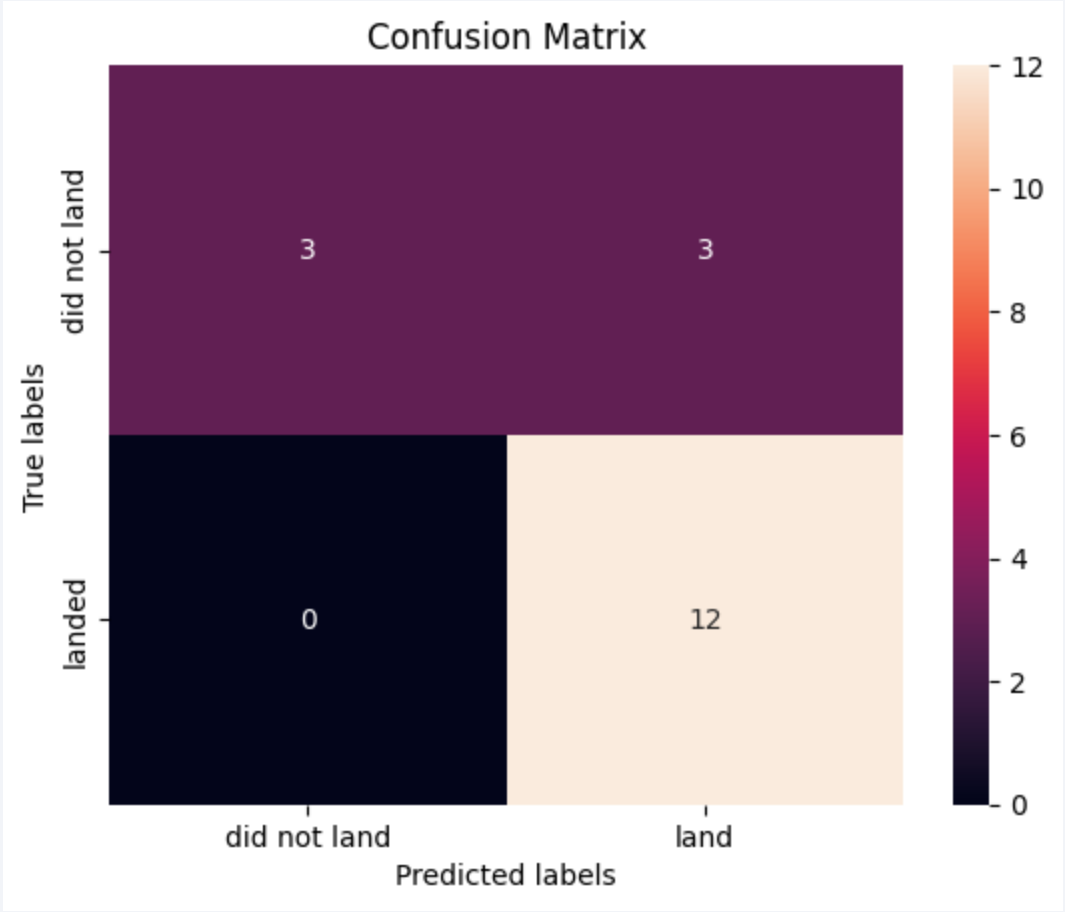
Classification Accuracy

- Four classification models were used to attempt to predict a successful landing: Logistic Regression, SVM, a Decision Tree, and KNN.
- Their relative training accuracies are plotted here.
- The Decision Tree has the highest relative accuracy, though by a small margin.
- With a larger dataset, I suspect a Random Forest classification algorithm may perform better here.



Confusion Matrix

- The best performing model's confusion matrix is attached.
- This shows that the model predicted 3 false positives, where model predicted the rocket would land when, in reality, it did not.
- This model predicted no false negatives and primarily predicted true positives.



Conclusions

Key Achievements

- Collected and cleaned launch data using SpaceX's API and web scraping from Wikipedia.
- Conducted detailed exploratory data analysis and built interactive visual tools to identify success patterns.

Predictive Modeling Outcomes

- Trained four classification models (Logistic Regression, SVM, Decision Tree, KNN) with ~83% predictive accuracy.
- Identified key features influencing first-stage recovery: launch site, payload mass, booster type, and orbit.

Business Impact

- Accurately predicting landing outcomes enables SpaceY to estimate launch costs and strategize for competitive pricing.
- Interactive dashboards empower non-technical stakeholders to explore and understand launch performance.

Future Directions

- Incorporate additional features (like weather data or mission type) to further refine predictive models.
- Explore advanced modeling techniques (ensemble methods, deep learning) to push prediction accuracy even higher.

Thank you!