IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Moustafa Elashwah
28 Aug 2025

Outline

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

Executive Summary

This project aims to leverage publicly available SpaceX data to build a machine learning model that predicts whether the Falcon 9 rocket's first stage will successfully land and be reused based on launch information.

The project applied the following techniques and methodologies:

- Data collection using APIs and web scraping
- Data wrangling and preprocessing
- Exploratory data analysis with Python and SQL
- Interactive mapping with Folium
- Dashboard development using Plotly Dash
- Predictive modeling and analysis

The project results include:

- A summary of key findings
- Descriptive statistics from the exploratory data analysis
- Data visualizations and interactive dashboards
- Insights from predictive analysis

Introduction

Background & Project Objectives

- The commercial space industry is highly competitive. SpaceX stands out by offering missions at a lower cost (~$62M vs. ~$165M for competitors) due to its ability to recover and reuse the Falcon 9's first stage.
- Predicting the likelihood of recovering and reusing the first stage is crucial for estimating mission costs—valuable both for SpaceX and for competitors bidding against them.

Project Goals:

- Collect and visualize SpaceX launch data in interactive dashboards.
- Analyze launch characteristics (payload mass, site, booster version) and their impact on reusability.
- Train predictive models using multiple machine learning algorithms.
- Evaluate and compare models to identify the most effective one.

Section 1

Methodology

Methodology

Data Collection:
- SpaceX REST API
- Web scraping from the SpaceX Wikipedia page

Data Wrangling:
- Filtering and sorting
- Handling missing values
- Creating a binary target variable for mission success/failure

Exploratory Data Analysis (EDA):
- Visualizations and SQL queries

Interactive Analytics:
- Folium maps and Plotly Dash dashboards

Predictive Modeling:
- Classification models to predict landing success
- Model training, evaluation, and comparison across multiple algorithms

Data Collection

Public SpaceX flight data was gathered through:

- REST API requests to the SpaceX API
- Web scraping from the SpaceX launch records on Wikipedia

Collected data included:

- Launch dates
- Launch site details
- Rocket booster specifications
- Payload mass
- First-stage recovery success/failure

Data Collection – SpaceX API

Launch data was retrieved through the SpaceX REST API, including details such as launch date, payload mass, booster version, launch site, and mission outcome.

[Data Collection Notebook Link]

Import “requests” library and send a get request to the SpaceX API → Decode the response as a JSON and turn it into a Pandas dataframe → Use custom functions and API call to populate missing data → Load collected data into Python dictionary → Convert dictionary into Pandas dataframe → Filter the dataframe to contain only data related to Falcon 9 launches → Handle missing values, replacing null payload masses with the mean of the column → Convert data frame to CSV and export

Data Collection - Scraping

Launch data was also gathered through web scraping, capturing details such as launch date, payload mass, booster version, launch site, and mission outcome.

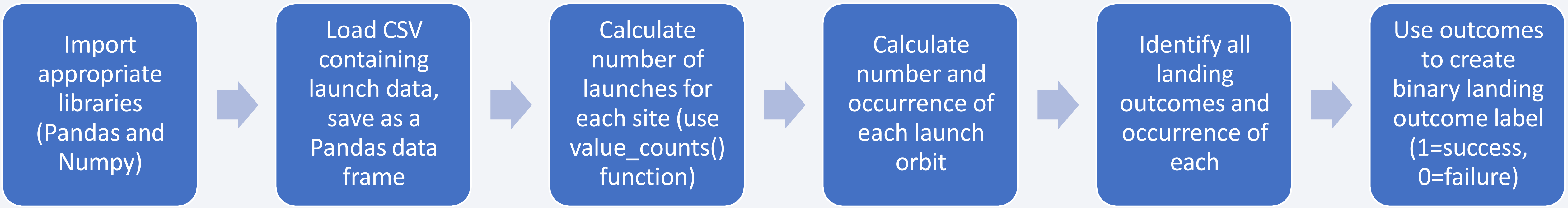
[webscraping Notebook Link]

1. Import “requests” library. Use get request on URL, save response as an object
2. Create a BeautifulSoup object using the response object
3. Use the “find_all” function to find all the HTML tables. ID desired table
4. Extract column headers by iterating through the table and apply custom function
5. Create dictionary using column headers as key.
6. Parse HTML table. Extract data and save to the dictionary.
7. Convert dictionary to Pandas data frame. Export as a CSV

Data Wrangling

- **Data Wrangling & Labeling**
- Prepared and cleaned collected data for analysis and modeling.
- Explored patterns and defined labels for supervised learning.
- Key tasks:
  - Counted launches per site and orbit type
  - Tallied landing outcomes
  - Created a binary landing label: **1 = successful landing**, **0 = unsuccessful**

Data Wrangling Notebook Link



EDA with Data Visualization

Created various plots to explore trends and patterns in the data:

- **Scatter Plots:** Examined relationships between variables, e.g.:
  - Flight Number vs Payload Mass (colored by launch outcome)
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
- **Bar Chart:** Compared success rates across different orbit types
- **Line Chart:** Showed annual success rates over time (2010–2020)

EDA Data Visualization Link

EDA with SQL

Performed SQL queries to explore launch data and uncover patterns.

- Key queries included:
  - Listing unique launch sites
  - Displaying first 5 records for sites starting with “CCA”
  - Calculating total and average payload mass for specific boosters
  - Identifying dates of first successful ground pad landings
  - Listing boosters that successfully landed on drone ships within 4000–6000 kg payload range
  - Counting successful vs unsuccessful missions
  - Finding boosters with maximum payloads
  - Extracting failed drone ship landings in 2015
  - Ranking landing outcomes between 2010–2017

SQL Notebook Link

Build an Interactive Map with Folium

- **Interactive Launch Map (Folium)**
- Built a geospatial map to visualize SpaceX launch sites and outcomes.
- **Features included:**
  - Circles and markers for each launch site with pop-up labels
  - Color-coded markers for launch results (**green = success, red = failure**)
  - Marker clusters to improve readability
  - Calculated distances from launch sites to nearby points of interest (highways, railroads, airports, cities)
  - Added PolyLines to show distances and connections between sites and key locations
  - MousePosition used to determine coordinates and calculate distances

Folium Notebook Link

Build a Dashboard with Plotly Dash

- Built a dynamic dashboard to explore launch data in real time.
- **Key features:**
- **Pie Chart:** Shows launch success percentages per site; updates for single or all sites
- **Scatter Chart:** Payload Mass vs. Launch Outcome by Booster version; highlights correlations and success rates
- **Range Slider:** Filters payload mass range on the scatter chart for interactive analysis

Dashboard Link

Predictive Analysis (Classification)

Create a NumPy array from the “Class” data (value want to predict)

→ Standardize , fit, and transform the data using StandardScaler. Save to “X”

→ Split the data into training and test data sets. (80% training, 20% test)

→ Train a logistic regression model, using GridSearchCV to find the best parameters.

→ Follow the same process to train SVM, Decision Tree, and KNN Models.

→ Calculate accuracy using .score() and create confusion matrix for each model.

→ Compare scores to determine the best performing model.

- Predictive Analysis Notebook

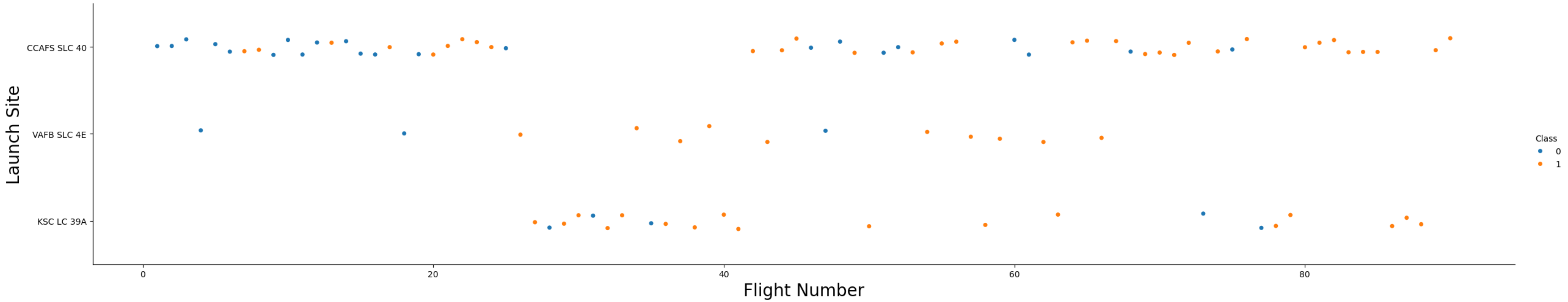
Results

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

Section 2

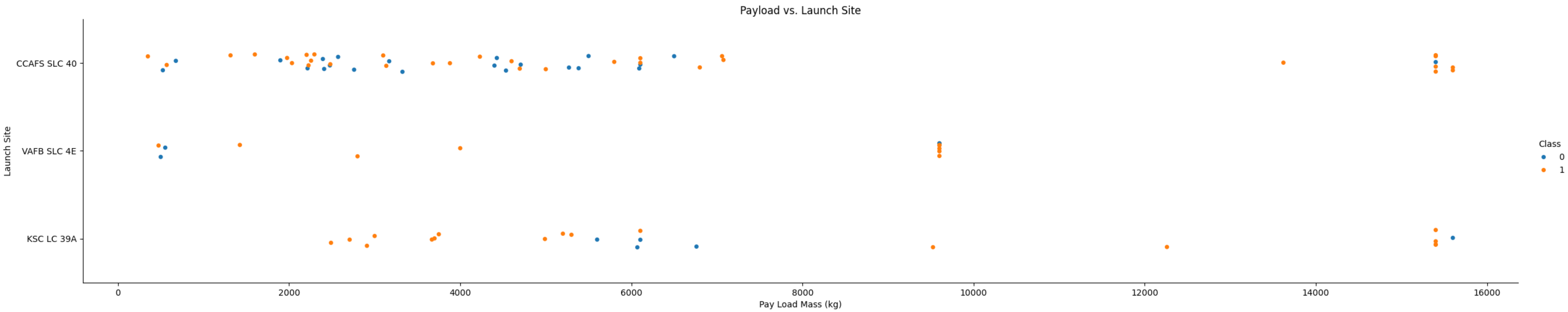
Insights drawn from EDA

Flight Number vs. Launch Site



- Flight numbers are on the x-axis, launch sites are on the y-axis, with blue data points indicating mission failure and orange data points indicating mission success.
- Site CCAFS SLC 40 had the highest number of launches, including 18 of the first 20 launches.
- Success rate improved over time, with early launches having a high failure rate, and later launches (#30 on) experiencing higher success rates.

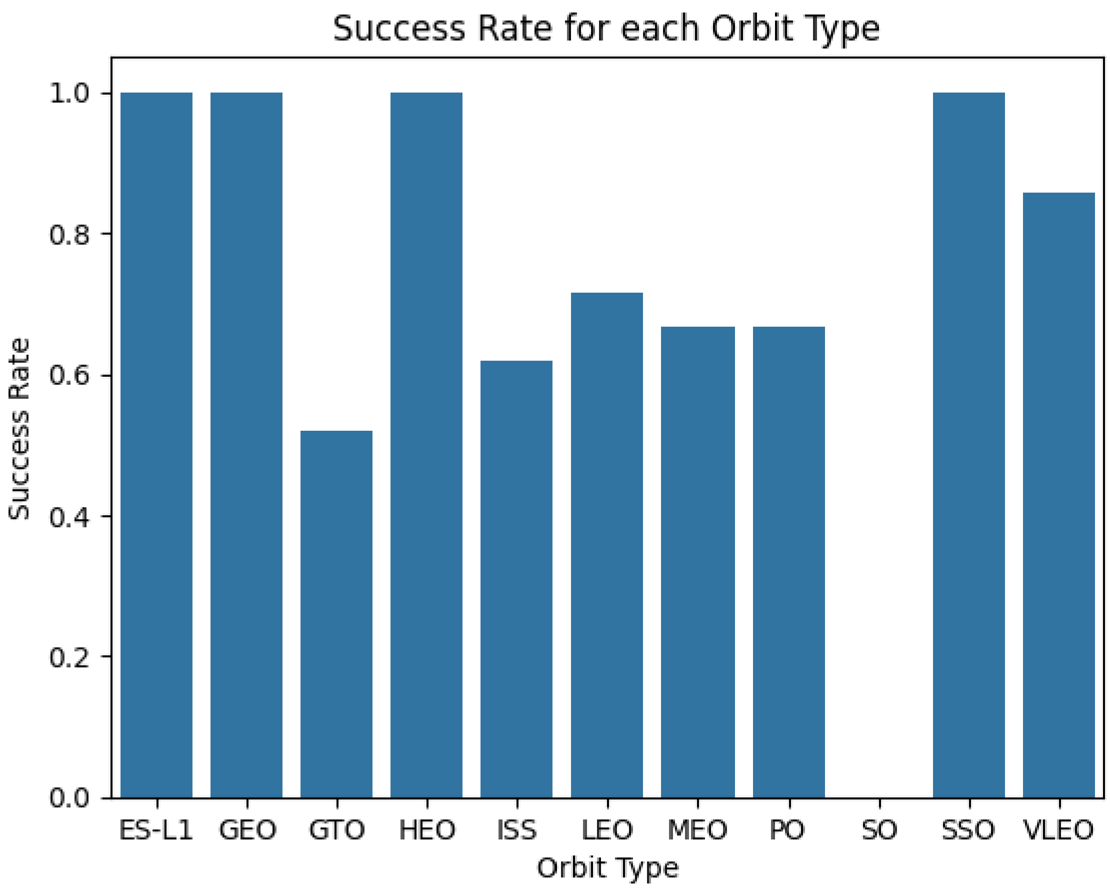
Payload vs. Launch Site



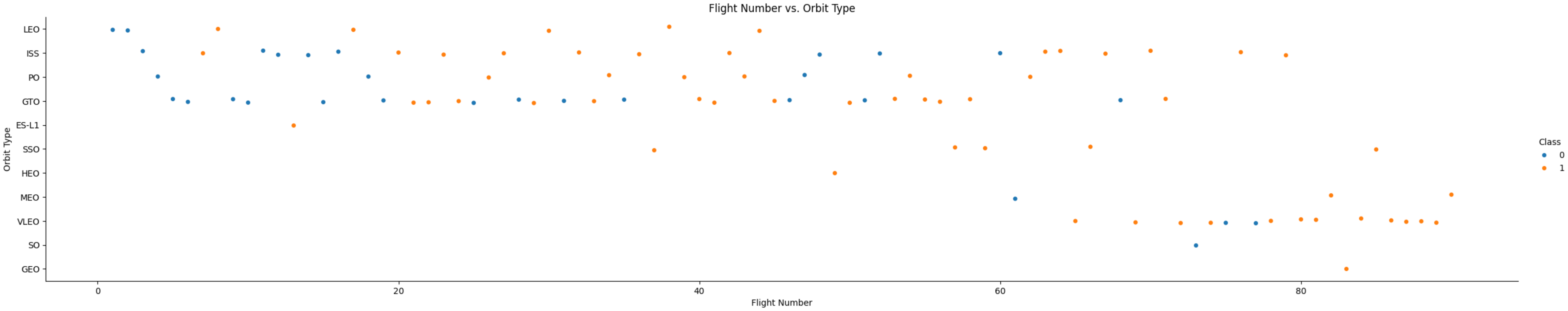
- Payload Mass (in kg) is on the x-axis, Launch Site is on the y-axis, with blue data points indicating failure, and orange data points representing success.
- The majority of the launches carried payloads less than 7,000 kg.
- Site VAFB SLC 4E did not launch a rocket with a payload greater than 10,000 kg.
- High payload launches (greater than 8,000 kg) experienced a high success rate.

Success Rate vs. Orbit Type

- Orbit type is the x-axis, success rate is on the y-axis.
- ES-L1, GEO, HEO, and SSO had the highest success rates at 100%.
- SO had the lowest success rate, at 0%.
- GTO, ISS, LEO, MEO, and PO all had success rates between 50% and 80%.

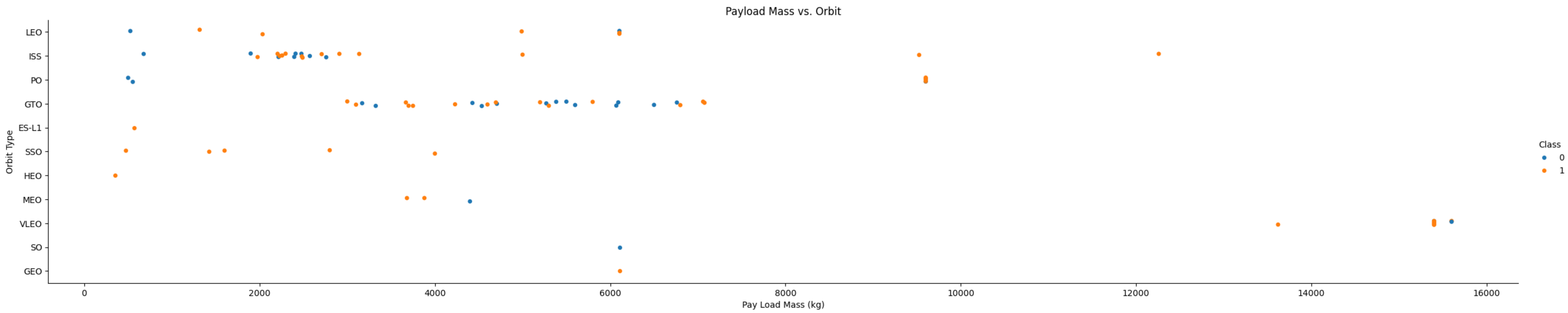


Flight Number vs. Orbit Type



- Flight number is on the x-axis, orbit type is on the y-axis, with blue data points indicating mission failure and orange data points indicating mission success.
- Majority of launches up to flight 55 had orbits of LEO, ISS, PO, or GTO.
- For LEO, success rate appears to improve over the launches, while GTO does not demonstrate a clear relationship.

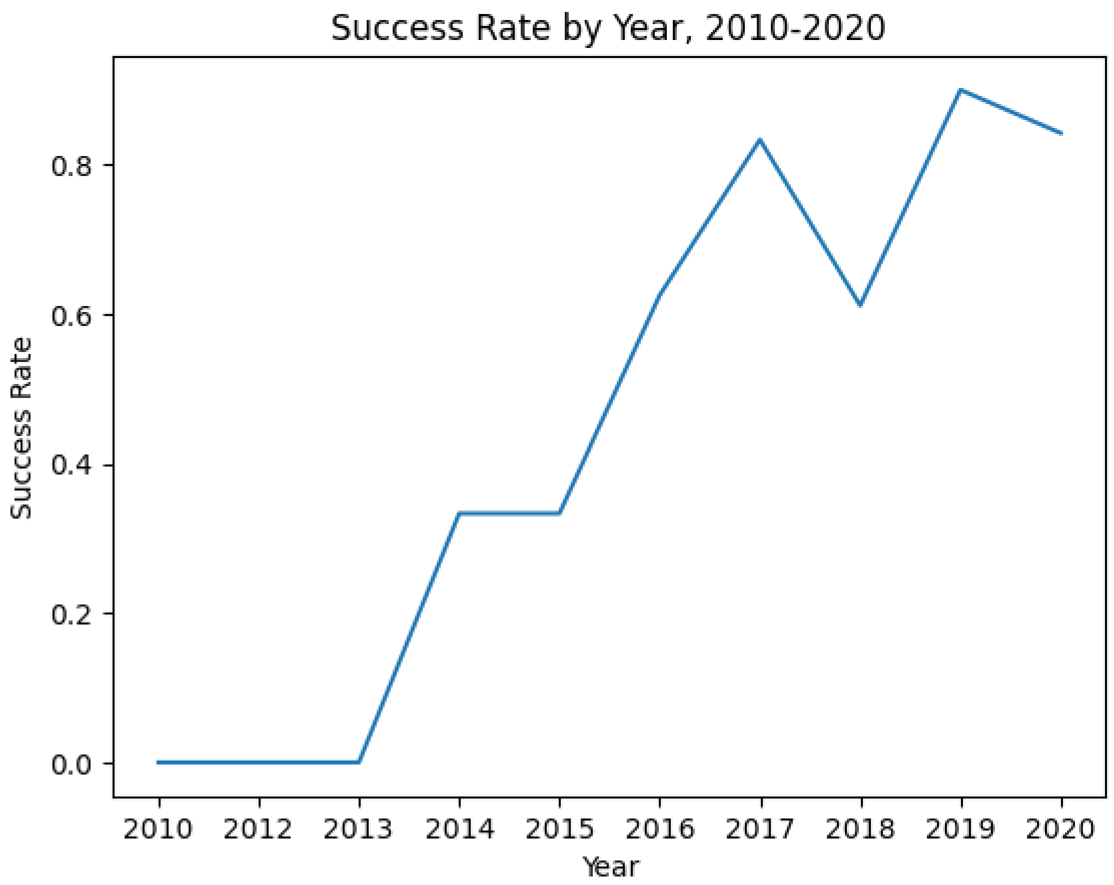
Payload vs. Orbit Type



- Payload Mass (in kg) is the x-axis, orbit type is the y-axis, with blue data points indicating mission failure and orange data points indicating success.
- Success rates for PO, ISS, and LEO increase as payload mass increases.
- GTO does not display any clear correlation between success and payload mass.

Launch Success Yearly Trend

- Year is the x-axis, success rate is the y-axis.
- Launches from 2010-2013 had a 0% success rate.
- Success rate improved between 2013-2020.



All Launch Site Names

- Task: Display all the launch sites.
- Query:
  - %sql Select DISTINCT(Launch_Site) from SPACEXTABLE
- “DISTINCT” displays the unique values from the “Launch_Site” column.
- Result:

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

Launch Site Names Begin with 'CCA'

- Task: Display 5 records where launch site begins with the string “CCA”
- Query:
  - %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' LIMIT 5
- Explanation:
  - like ‘CCA%’ selects all records where the launch site starts with CCA.
-

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

- Task: Display the total payload mass carried by boosters launched by NASA (CRS)
- Query:
  - %sql select SUM(PAYLOAD_MASS__KG_) AS 'Total_Payload_Mass_KG' from SPACEXTABLE where Customer = 'NASA (CRS)’
- Explanation:
  - The WHERE clause filters for records with a customer value equal to “NASA (CRS)”
  - SUM(PAYLOAD_MASS_KG_) displays the sum of this column for the filtered records.
- Result:

Total_Payload_Mass_KG
45596

Average Payload Mass by F9 v1.1

- Task: Display average payload mass carried by booster version F9 v1.1
- Query:
  - %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1’
- Explanation:
  - WHERE clause filters records to display records matching the specified booster version.
  - AVG(PAYLOAD_MASS__KG_) calculates the average value for payload mass column of the filtered records.
- Result:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Task: List the date when the first successful landing outcome in ground pad was achieved.
- Query:
    - %sql select MIN(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)’
- Explanation:
    - WHERE clause limits the query to records where landing outcome equals the specified value.
    - MIN(Date) selects the lowest/earliest date value.
- Result:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Task: List the name of the boosters which have success in drone ship landing and have a payload mass greater than 4000 but less than 6000.
- Query:
  - %sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ between 4000 and 6000
- Explanation:
  - WHERE clause sets payload mass range and filters for successful drone ship landings.
- Result:

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

Total Number of Successful and Failure Mission Outcomes

- Task: List the total number of successful and failure mission outcomes.
- Query:
  - %sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
- Explanation:
  - GROUP BY clause groups values by the unique values in the column.
  - Count(*) displays the total number of records in each group.
- Result:

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

Boosters Carried Maximum Payload

- Task: List the names of the booster versions which have carried the maximum payload mass.
- Query:
  - %sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
- Explanation: Used a sub-query since WHERE clauses cannot contain aggregate functions.
- Result:

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

- Task: List the records from 2015 that failed drone ship landings.
- Query:
  - %sql select substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE WHERE substr(Date, 0,5) = '2015' AND Landing_Outcome = 'Failure (drone ship)’
- Explanation:
  - WHERE clause sets year and outcome parameters.
  - SELECT clause specifies which values to display.
- Result:

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

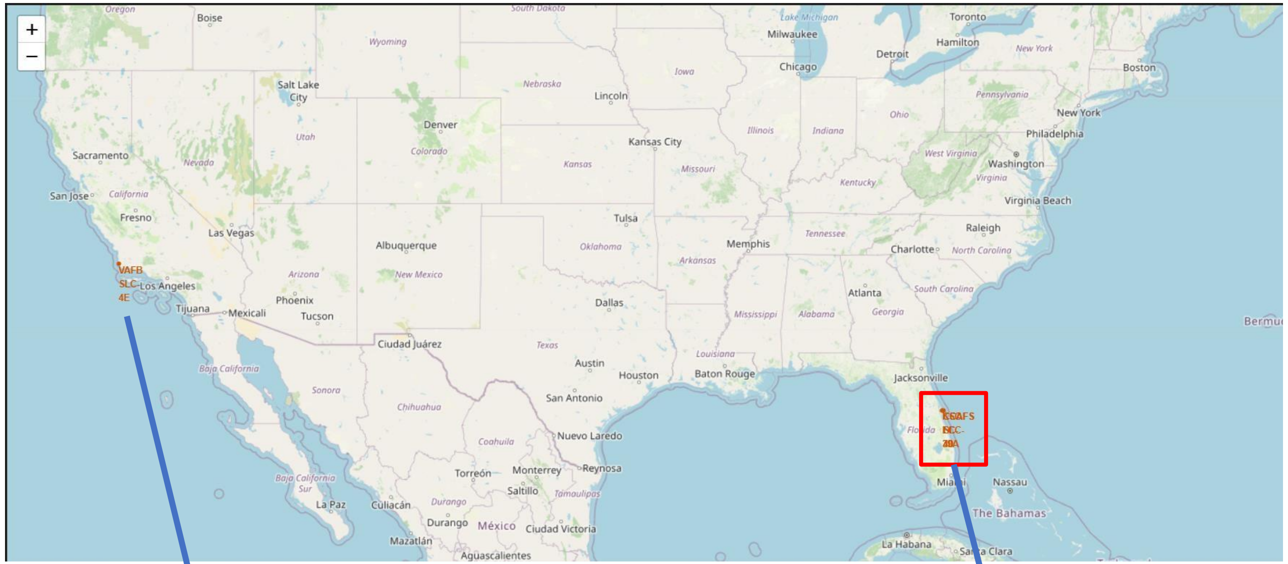
- Task: Rank the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order.
- Query:
  - %sql select Landing_Outcome, count(*) FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' and '2017-03-20' Group By Landing_Outcome Order By count(*) DESC
- Explanation:
  - GROUP BY clause groups records into the various landing outcomes.
  - HAVING clause sets the date range for the records.
  - DESC orders the results from largest to smallest.
- Result:

Landing_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

Map of All SpaceX Falcon 9 Launches

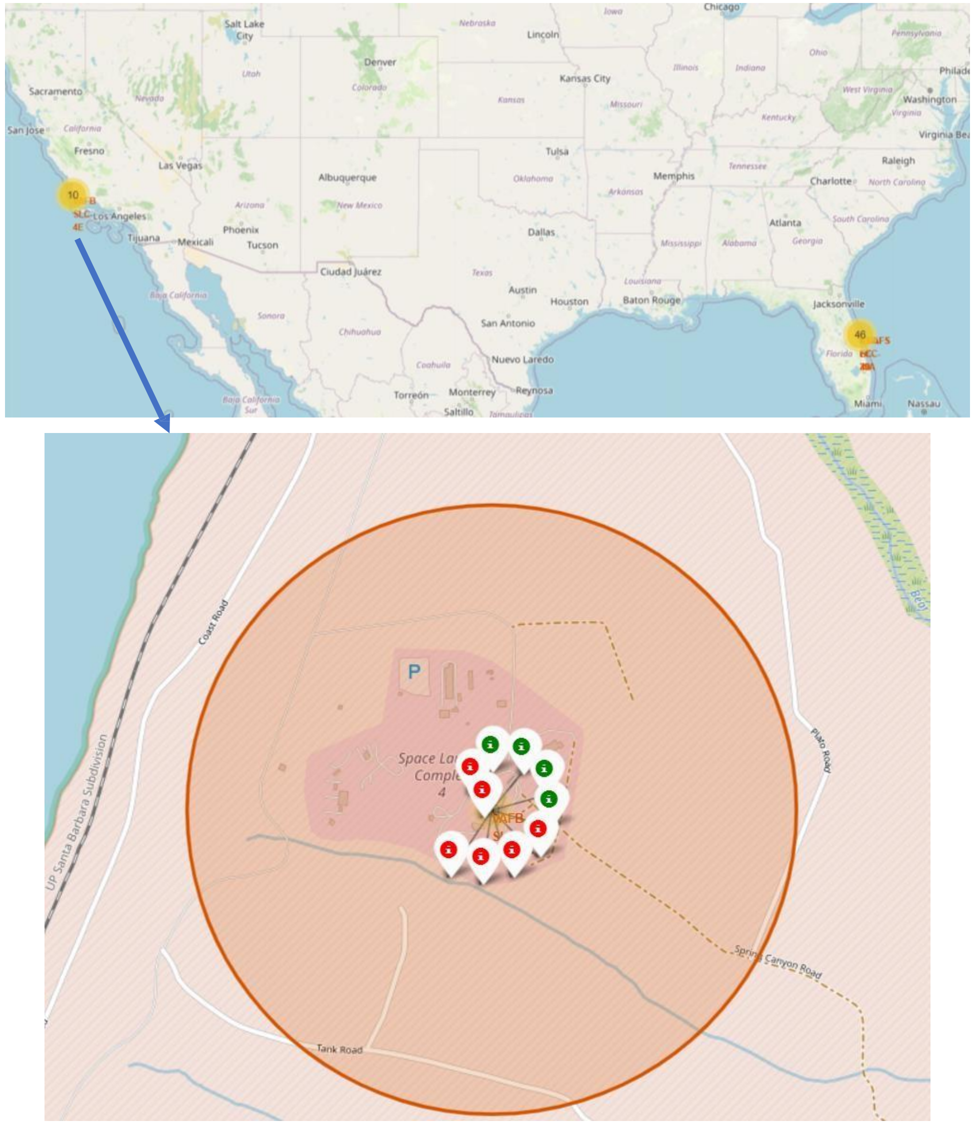


This map shows the location of the four launch sites. The bottom two images are zoomed in to show more detail.

Sites are denoted by a Circle with a Marker as the text label.

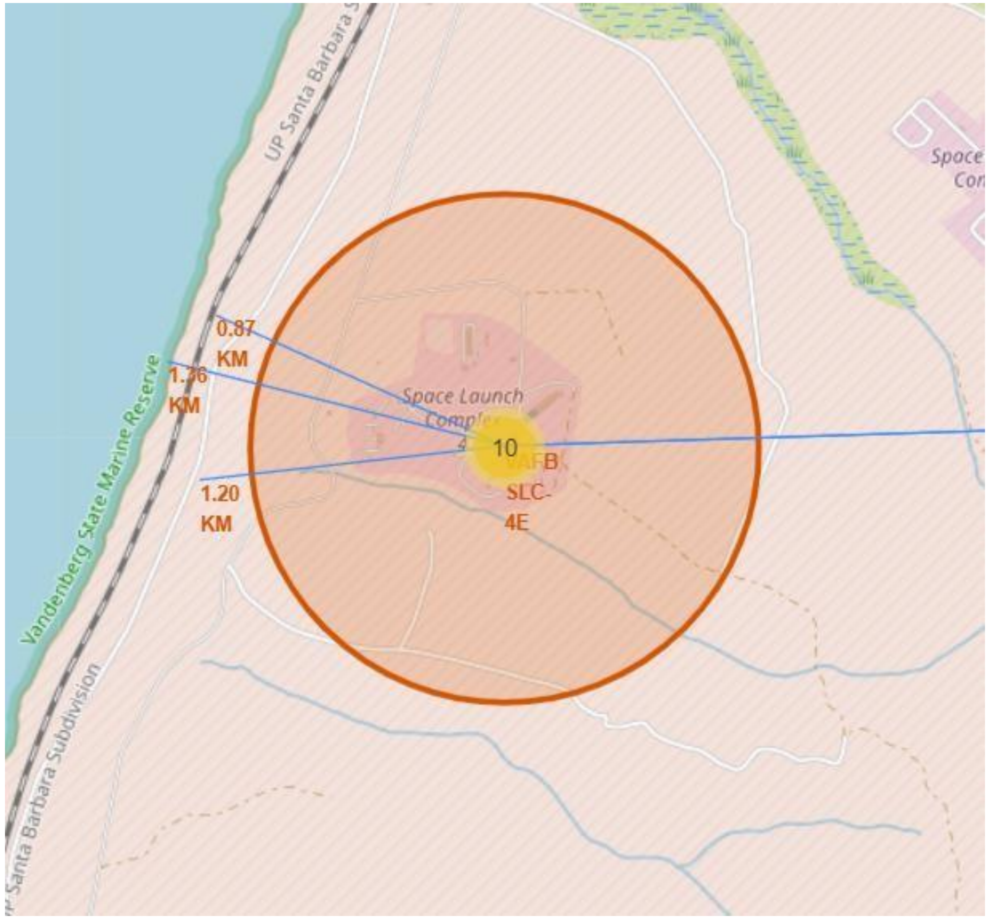
All launch sites are in the southern portion of the United States and are close to the coast.

Launch Outcomes By Site



- Added Marker Clusters to each launch site to indicate the number of launches at each site.
- The top map illustrates the small scale view. Yellow circles represent the clusters, the number showing the number of launches.
- The bottom map shows a zoomed in view of the VAFB SLC 4E launch site. Markers in the cluster are assigned a color:
  - Red – Failed landing
  - Green – Successful landing

Launch Site Proximity to Points of Interest



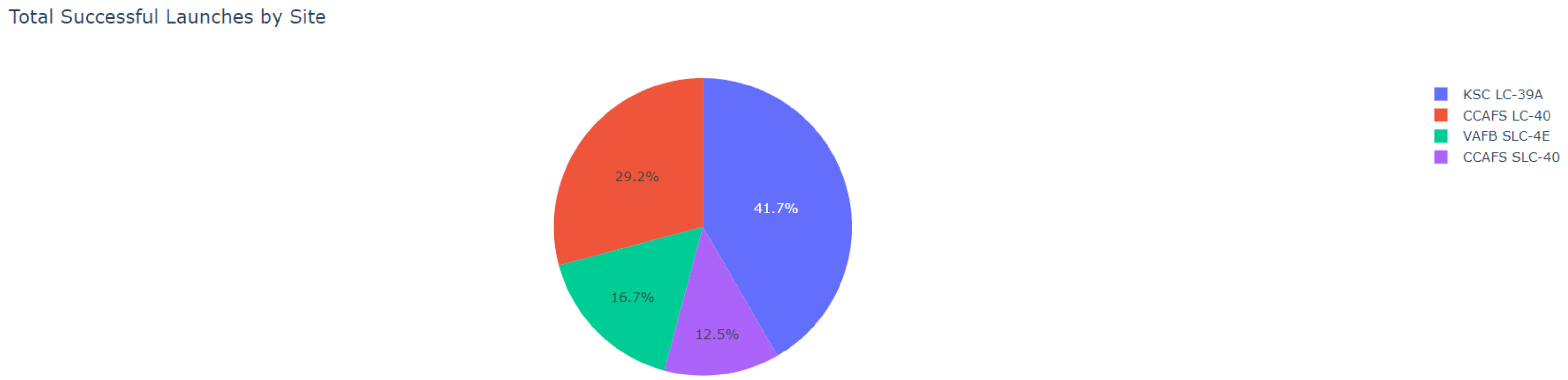
- This map shows the distance from launch site VAFB SLC 4E to various points of interest.
- Distances are represented by PolyLines, with markers showing the distance each line represents.
- VAFB is:
  - 0.87 km from the nearest railroad
  - 1.36 km from the coast
  - 1.2 km from the nearest highway
  - 14 km from the nearest city/airport
- All launch sites are near the coast to launch rockets over the water and are near a major transportation route (highway/railroad)

Section 4

Build a Dashboard with Plotly Dash

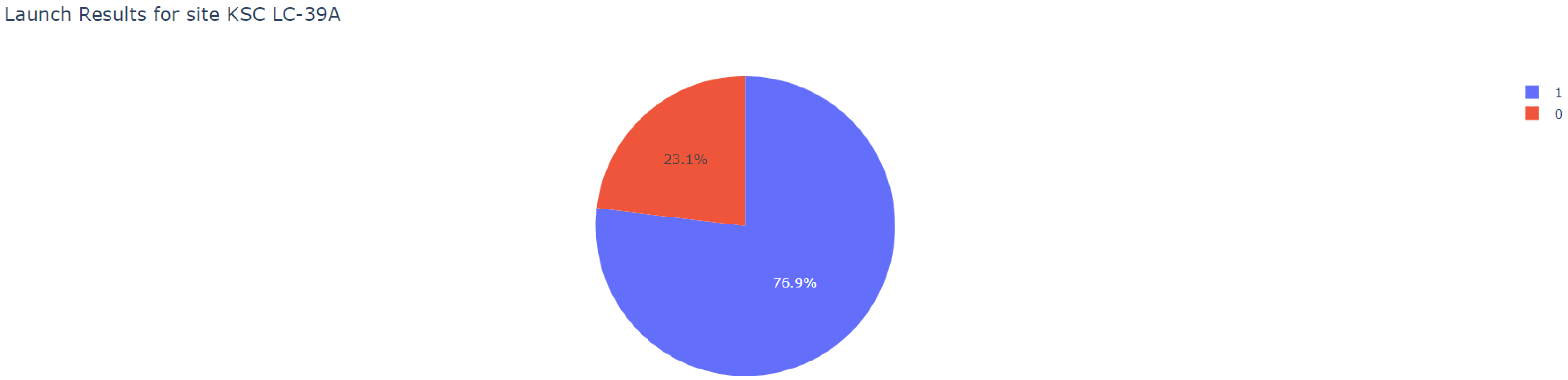
Total Successful Launches, By Site

All Sites



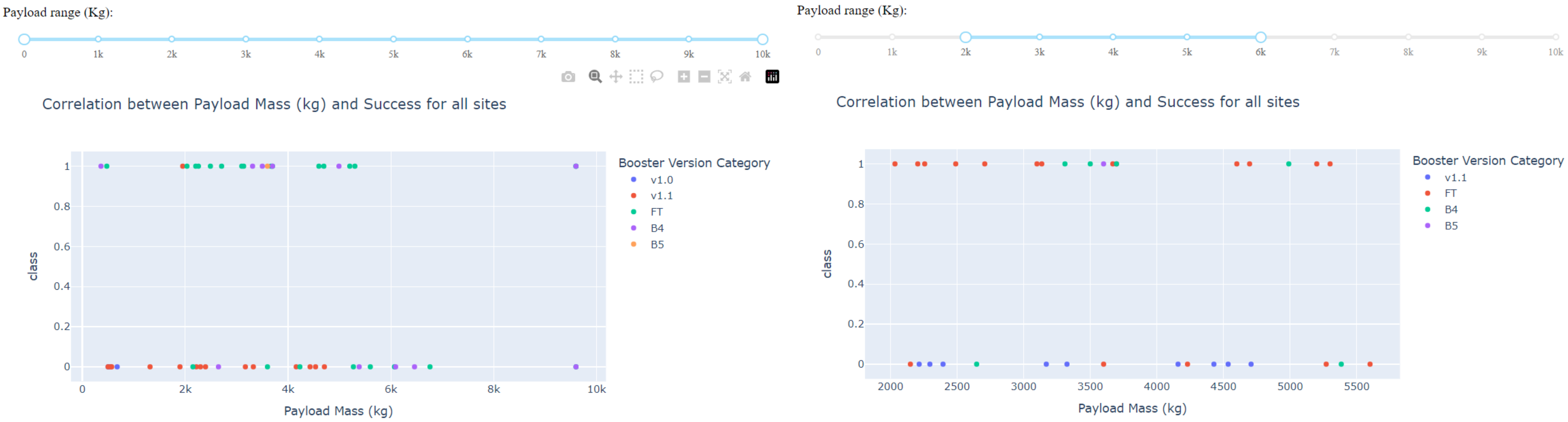
- Pie chart showing total launch successes among all sites.
- KSC LC-39A has the highest percent of successes at 41.7%
- CCAFS SLC-40 has the lowest percent of successes at 12.5%

Launch Results for KSC LC-39A



- KSC LC-39A – Launch site with the highest number of successful launches.
- Site has a success rate of 76.9%
- 23.1% of launches at this site failed.

Payload Mass vs. Success Rate, All Sites

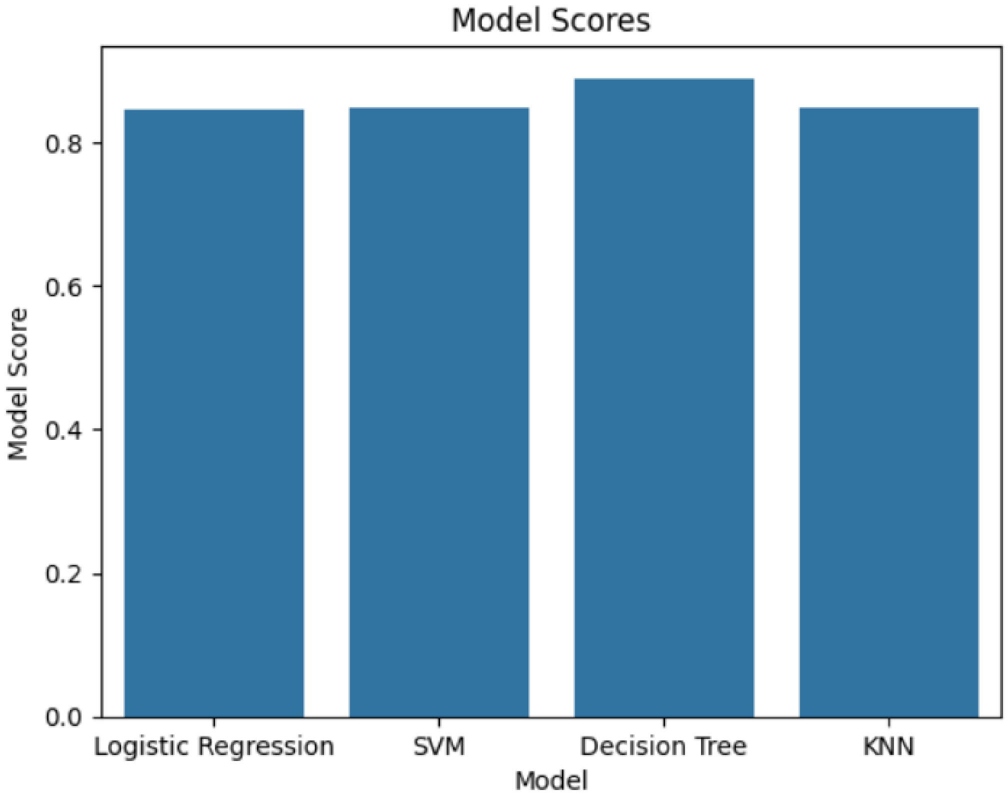


- The left plot shows the launch outcome (y-axis) for all payload masses (x-axis).
- Most of the successful launches occur when payload mass is between 2000 kg and 5500 kg, shown by the plot on the right.

Section 5

Predictive Analysis (Classification)

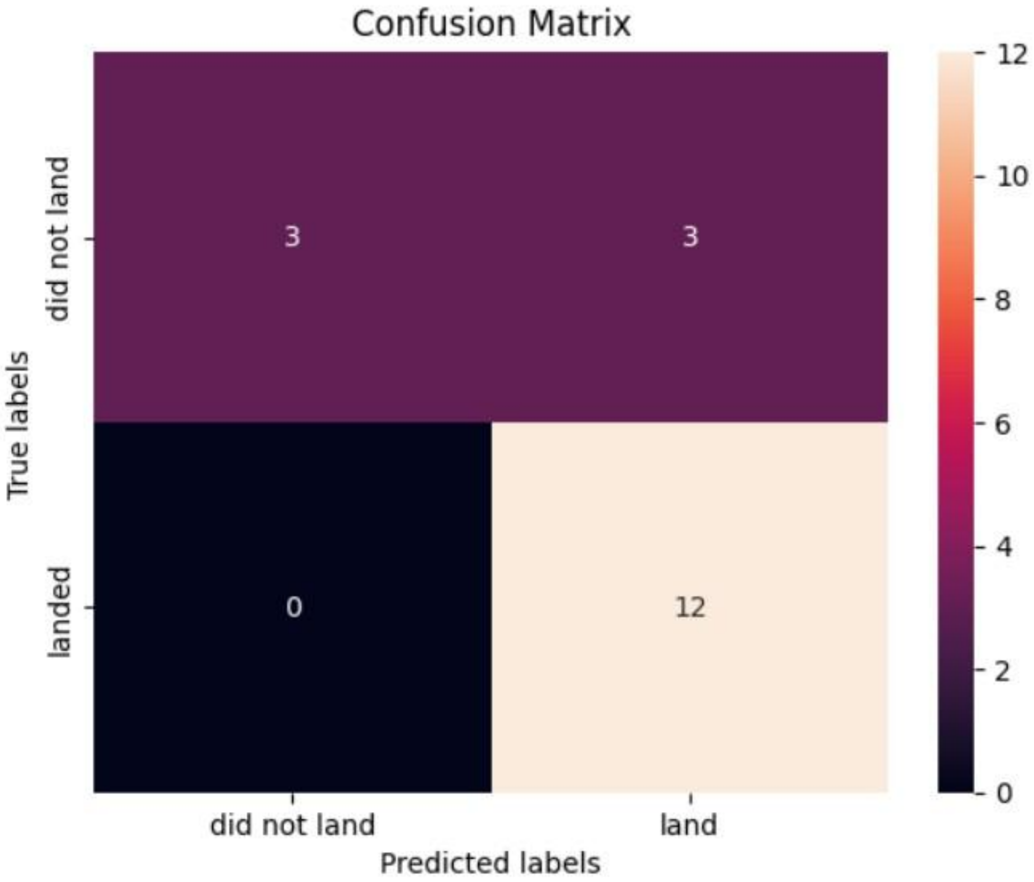
Classification Accuracy



	Model	Model Score	Model Test Data Score
0	Logistic Regression	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Decision Tree	0.889286	0.833333
3	KNN	0.848214	0.833333

- The Decision Tree Classification Model scored the best of the four models.
- All four models have similar classification scores.
  - Highest = Decision Tree (0.889)
  - Lowest = Logistic Regression (0.846)
- All models have the same accuracy score on the test data set (0.833).
- As new data becomes available for training, one model may appear as the definitive best.

Confusion Matrix



- All confusion matrixes were the same.
- Models predicted the outcome of 18 launches.
  - Accurately predicted 15 of 18 outcomes. (83.3%)
  - 3 of the predicted successes failed. (16.7%)
- These are Type 1 Errors (false positives).
  - Type 1 Error are less desirable than Type 2.
- Type 1 Errors can result in underestimating the actual cost of a launch, as fewer rockets can successfully be reused than initially predicted.

Conclusions

- Findings from Exploratory Data Analysis (EDA):
  - As more rockets are launched, success rate improves (flight number and success rate positively correlated).
  - ES-L1, GEO, HEO, and SSO orbits had the highest success rates (100%).
  - Success rates improved from 2013-2020, from 0% to ~80%.
- Findings from Proximities Analysis:
  - Launch sites are in the southern United States, as near the equator as practical.
  - Launch sites are near the coast and a major highway or railroad.
- From the Interactive Dashboard:
  - KSC LC-39A had the most successful launches of all the sites.
  - Most successful launches had a payload mass between 2,000 kg and 5,500 kg.
- From Predictive Analysis:
  - Decision Tree Classification scored the best, but all four models performed similarly well.
  - All models experienced Type I errors, which is the less desirable error and can result in underestimate costs.
  - As new data is available, using it to train/test the data should improve results.

Thank you!