IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chris Popiel
04 March 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Public Content

Executive Summary

The purpose of this project is to use public data from SpaceX to train a machine learning model. The model will take launch information and predict if SpaceX will be able to successfully land and reuse the first stage from the Falcon 9 rocket.

The techniques and methodologies used during this project include:

- Data Collection via API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis with Python and SQL
- Constructing interactive maps with Folium
- Dashboard Construction with Plotly Dash
- Predictive Analysis

- Summary of results discussed
  - Descriptive statistics from Exploratory Data Analysis
  - Data visualizations and dashboards
  - Predictive analysis findings

Public Content

Introduction

- There is heavy competition in the commercial space industry, and SpaceX has distinguished itself from its rivals by offering relatively less expensive space missions, ~$62M, compared to its competitors, ~$165M. SpaceX is able to do so because of their ability to recover and reuse the first stage of their rockets.
- The ability to predict the likelihood of recovering and reusing the first stage of a launch is invaluable to determining the overall cost of a mission. This information is important for SpaceX, and for any companies that might consider bidding against them.
- This project seeks to:
  - Gather information about SpaceX launches and present the data in dashboards.
  - Examine launch characteristics, such as payload mass, launch site, and booster version, and attempt to quantify their impact on reusing the first stage of the rocket.
  - Utilize multiple algorithms to train models to predict the likelihood of recovering the first stage.
  - Evaluate the models to determine the most effective one.

Public Content

Section 1

Methodology

Public Content

Methodology

- Data was collected in two ways:
  - Using the SpaceX REST API
  - Web scraping data with Python from SpaceX Wikipedia site
- Perform data wrangling
  - Filtered and sorted data
  - Addressed missing values
  - Create binary label to indicate mission success or failure
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, train, and evaluate models using various algorithms

Public Content

Data Collection

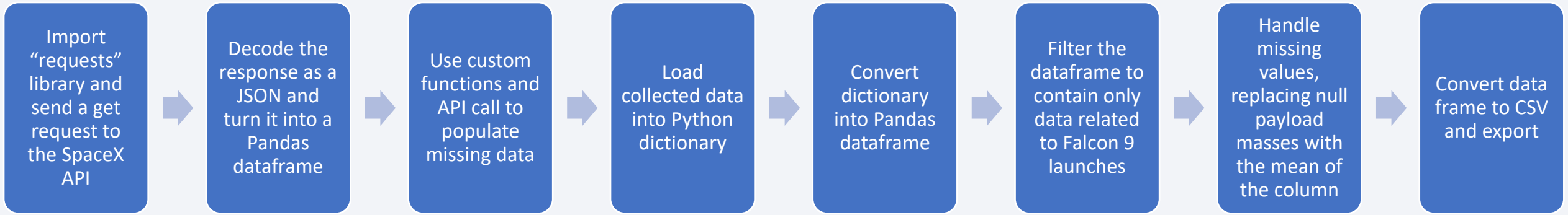
- Public data related to SpaceX flights were collected via two methods:
  - Get request to the SpaceX Representational State Transfer (REST) Application Programming Interface (API)
  - Web scrapping from Wikipedia page of SpaceX launch records
- Data collected included:
  - Launch Dates
  - Launch Site Information
  - Rocket Booster Information
  - Payload Mass
  - Success/Failure of recovering first stage

Public Content

Data Collection – SpaceX REST API

- Data collected via API followed the below process.
- Launch data collected included date, payload size, rocket booster version, launch site, and launch outcome.

SpaceX REST API process



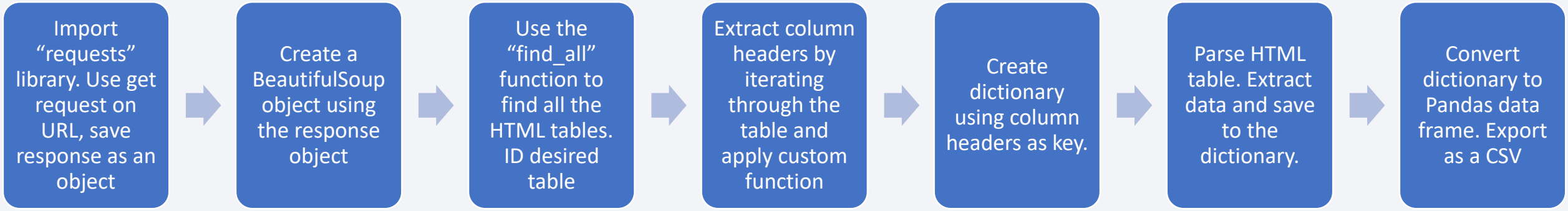
GitHub Link: API

Public Content

Data Collection – Scraping

- Data collected via web scraping followed the below process.
- Launch data collected included date, payload size, rocket booster version, launch site, and launch outcome.

SpaceX REST API process



GitHub Link – Web Scraping

Public Content

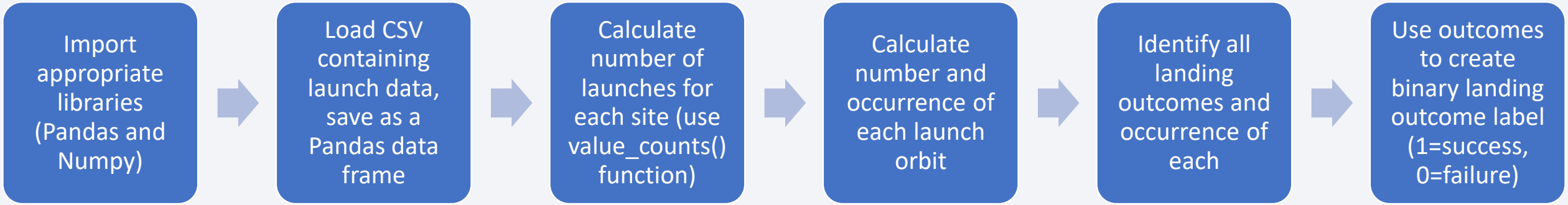
Data Wrangling

- Conducted data wrangling on the collected data.
- The purpose of data wrangling was to perform initial exploratory data analysis (EDA) and identify potential patterns in the data and define labels for training supervised learning models.
- Tasks in this step included:
  - Calculating the number of launches at each site
  - Calculating the number and occurrence of each launch orbit
  - Calculating the number of each landing outcome
  - Creating a binary landing outcome label
- For the landing outcome label, “1” represents the first stage booster successfully landed, and “0” represents the booster was unsuccessful in landing.

Public Content

Data Wrangling - Continued

Data Wrangling Process



GitHub Link – Data Wrangling

Public Content

EDA with Data Visualization

- As part of the EDA process, several plots were created to examine trends in the data.
- **Scatter Plot**: Show the relationship/correlation between two variables. Used to identify patterns. The following scatter plots were created:
  - Flight Number vs Payload Mass, with color indicating launch outcome
  - Flight Number vs Launch Site Location, with color indicating launch outcome
  - Payload Mass vs Launch Site Location, with color indicating launch outcome
  - Flight Number vs Orbit Type, with color indicating launch outcome
  - Payload Mass vs Orbit Type, with color indicating launch outcome
- **Bar Chart**: Used to compare values among discrete categories. The bar chart created for this analysis illustrated success rate for each launch orbit type.
- **Line Chart**: Typically used to show time series trends. The line chart created for this analysis illustrated annual success rate over time (from 2010-2020)

GitHub Link – EDA with Visualization

Public Content

EDA with SQL

- Used SQL to conduct additional EDA on launch data, performing various queries to better understand the data and identify any trends or patterns.
- The following queries were performed:
  - Display the names of the unique launch sites
  - Display 5 records where launch sites begin with the string “CCA”
  - Display the total payload mass carred by boosters launched for NASA (CRS)
  - Display the average payload mass carried by F9 v1.1 boosters
  - List the date when the first successful landing outcome on a ground pad was achieved
  - List the names of the boosters which landed successfully on a drone ship and have a payload mass between 4000 kg and 6000 kg
  - List the total number of successful and unsuccessful mission outcomes
  - List the names of the booster versions which carried the maximum payload mass
  - List records that failed landings on drone ships in 2015
  - Rank the count of landing outcomes between 06/04/2010 and 03/20/2017 in descending order

GitHub Link – EDA with SQL

Public Content

Build an Interactive Map with Folium

- Built an interactive map using the Folium library to illustrate geospatial data related to the launches.
- First, created to a map to show all launch sites.
  - Added circles to denote the location of each launch site, with a popup label displaying the site name.
  - Added markers to display the name of the launch site by each circle.
- Next, indicated the result of the launches at each site.
  - Added markers for each launch and added color to indicate success (green) or failure (red).
  - Created marker clusters at each site to improve readability.
- Last, calculated distance from each launch site to nearby points of interest (highway, railroad, airport, etc.).
  - Added MousePosition to determine coordinates and wrote function to calculate distances between coordinates.
  - Added a PolyLine between site CCAFS SLC-40 and the coastline, with distance as the label.
  - Added a PolyLine with distance between site VAFB SLC-4E and the nearest railroad.
  - Added a PloyLine with distance between site VAFB SLC-4E and the nearest city.

GitHub Link - Folium

Public Content

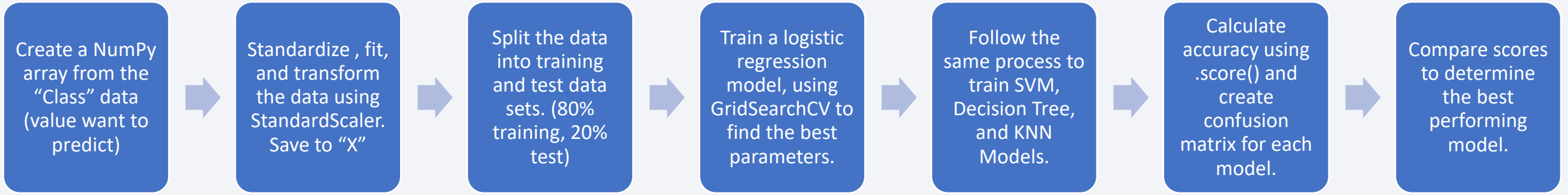
Build a Dashboard with Plotly Dash

- Used Plotly Dash to build an interactive dashboard, allowing users to adjust parameters and see updated charts in real time.
- Created a Pie Chart with a dropdown menu listing the launch sites.
  - When all launch sites selected, pie chart displays the percent of successful launches at each site.
  - When a single launch site selected, pie chart displays number of successes and failures at that site.
  - This is a useful visualization for identifying which site experienced the most successful launches.
- Created a Scatter Chart of Payload Mass vs. Launch Outcomes for each Booster version.
  - Displays any correlation between payload mass and success rates.
  - Coloring points by Booster version provides additional information which Boosters have the highest success rates.
  - Created range slider for Payload Mass, allowing the user to set a range for the x-axis on the chart.

GitHub Link – Dash App

Public Content

Predictive Analysis (Classification)



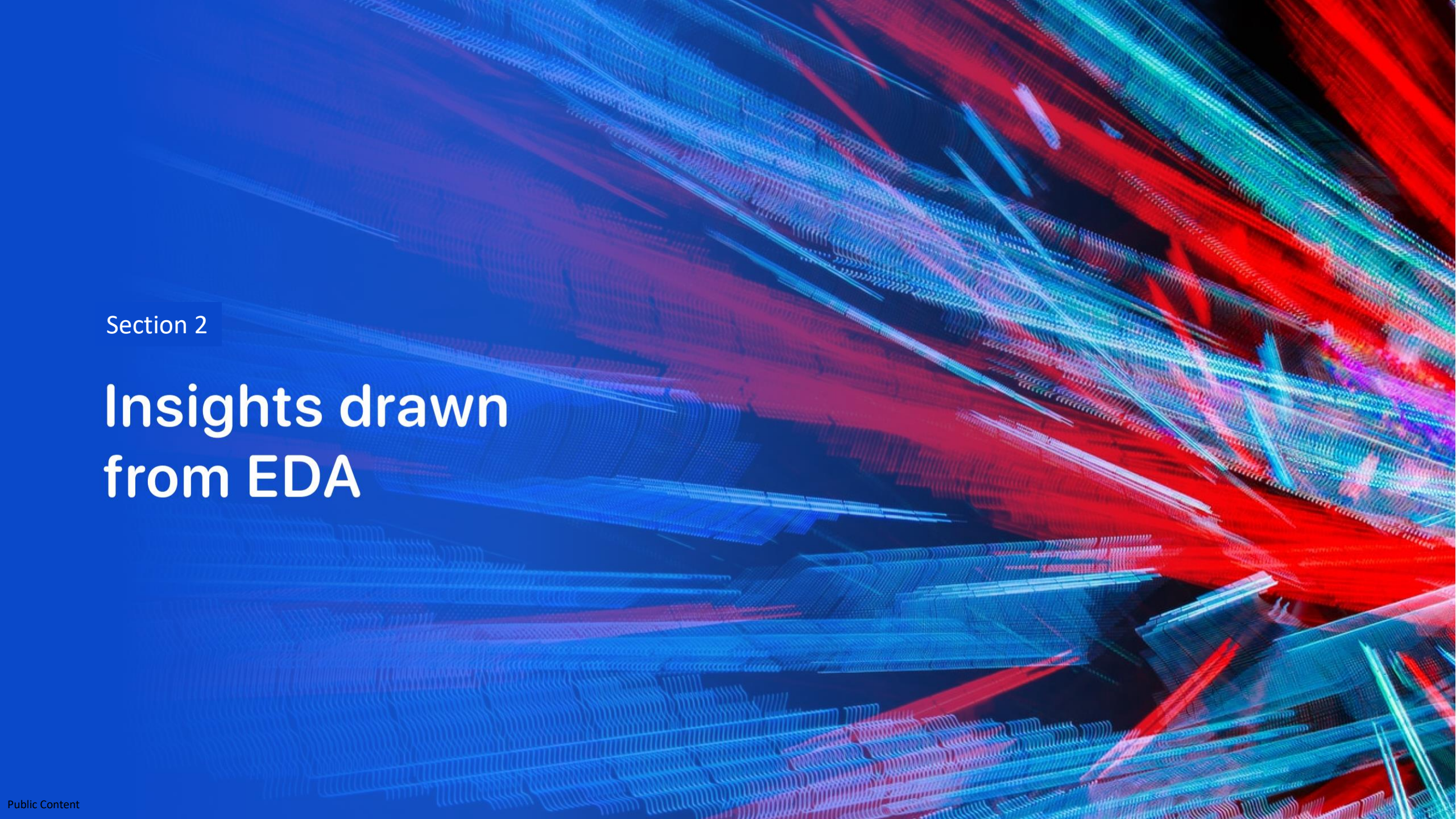
GitHub Link – Predictive Analysis

Public Content

Results

- The following slides will discuss the results of the analysis, including:
  - Exploratory data analysis findings.
  - Screenshots illustrating the interactive analytics dashboard.
  - Predictive analysis results and model comparison.

Public Content

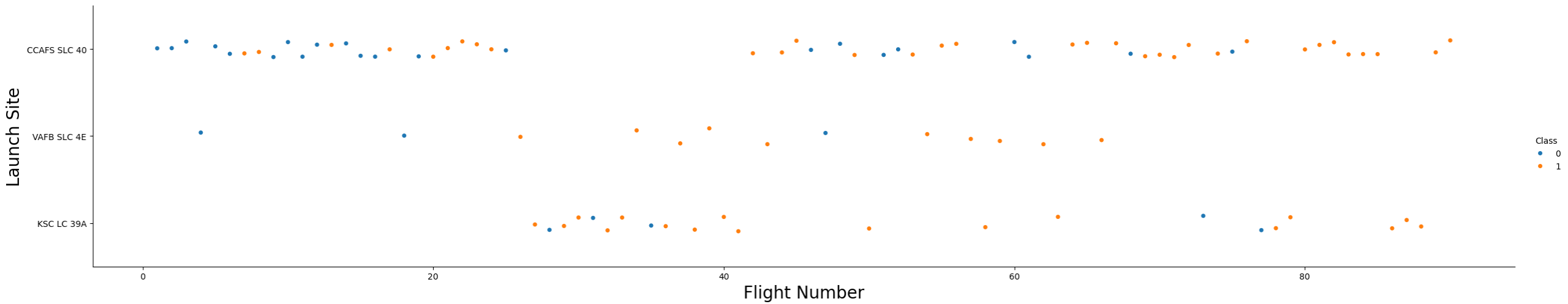


Section 2

Insights drawn from EDA

Public Content

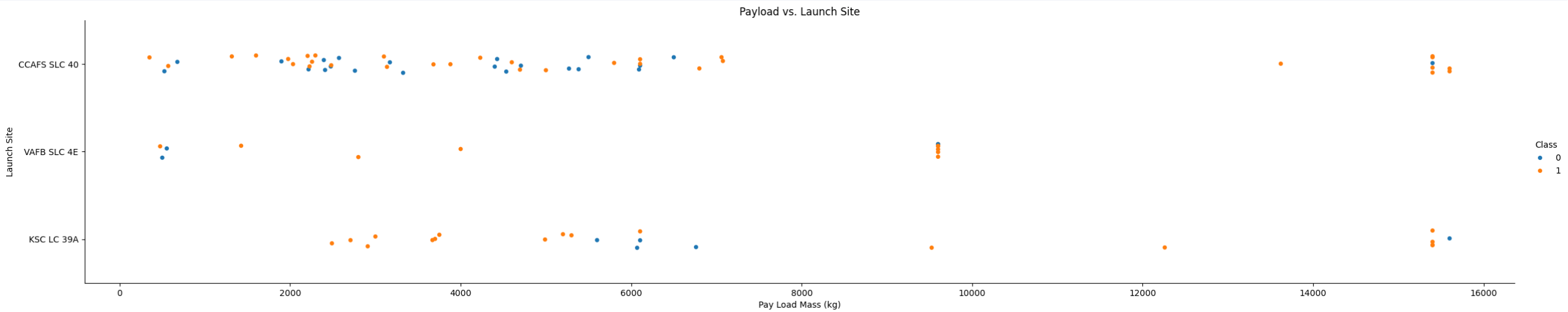
Flight Number vs. Launch Site



- 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.

Public Content

Payload vs. Launch Site

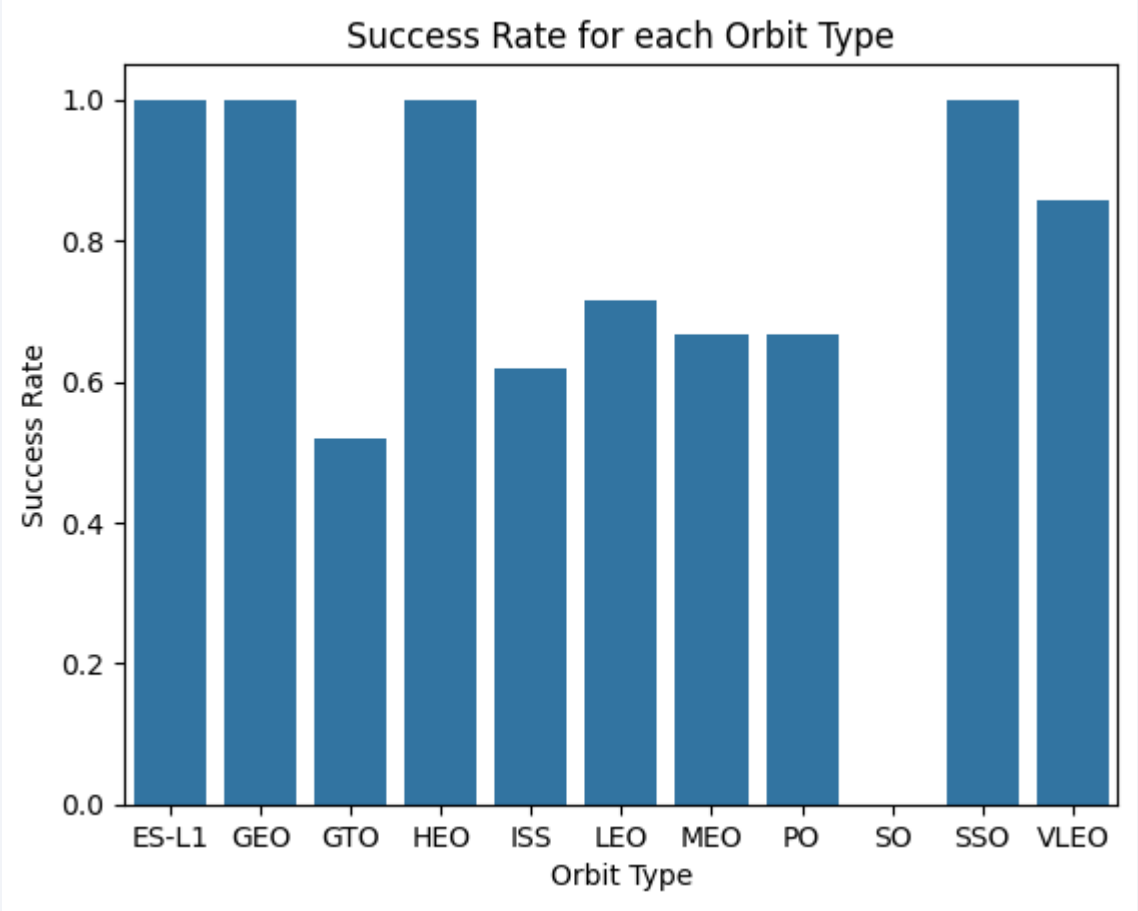


- Payload Mass (in kg) is on the x-axis, Launch Site is on the y-axis, with blue data points indicating failure, and orange data points representing success.
- The majority of the launches carried payloads less than 7,000 kg.
- Site VAFB SLC 4E did not launch a rocket with a payload greater than 10,000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

Public Content

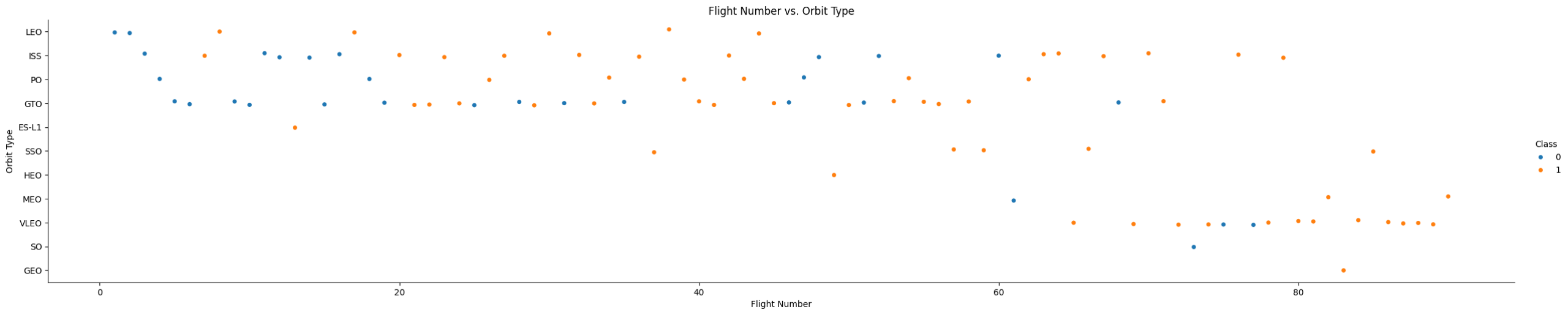
Success Rate vs. Orbit Type

- 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%.



Public Content

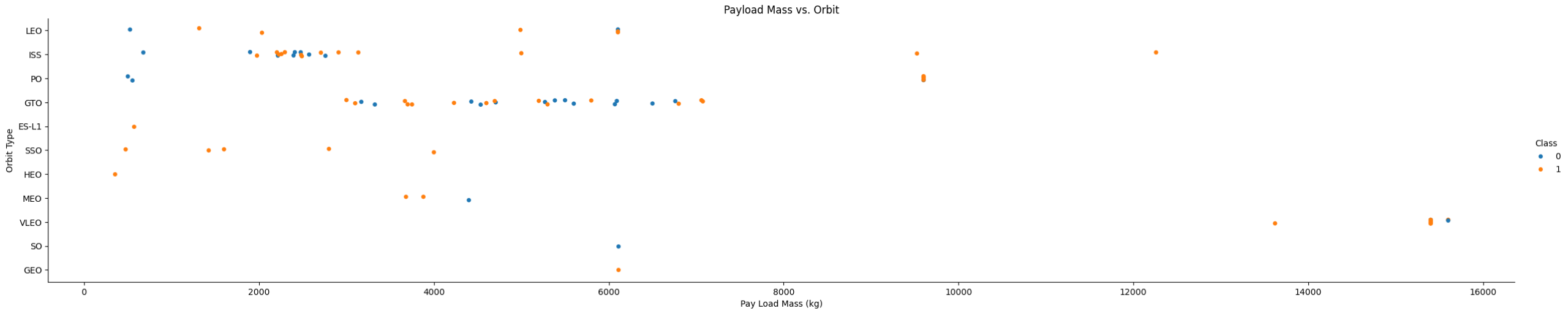
Flight Number vs. Orbit Type



- 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.

Public Content

Payload vs. Orbit Type

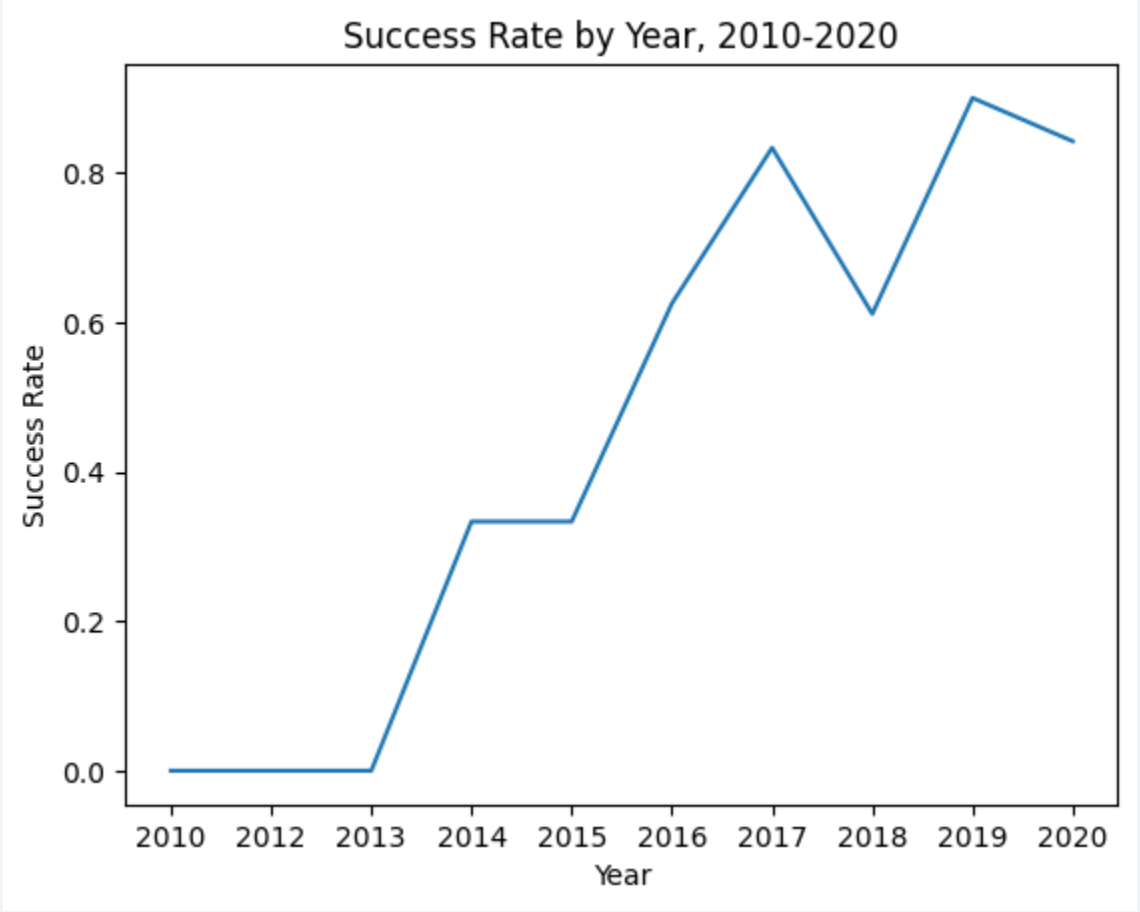


- Payload Mass (in kg) is the x-axis, orbit type is the y-axis, with blue data points indicating mission failure and orange data points indicating success.
- Success rates for PO, ISS, and LEO increase as payload mass increases.
- GTO does not display any clear correlation between success and payload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

Public Content

Launch Success Yearly Trend

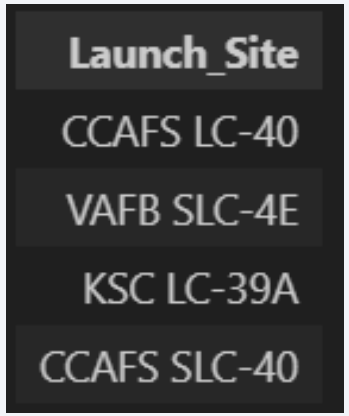
- Year is the x-axis, success rate is the y-axis.
- Launches from 2010-2013 had a 0% success rate.
- Success rate improved between 2013-2020.



Public Content

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:



Public Content

Launch Site Names Begin with 'CCA'

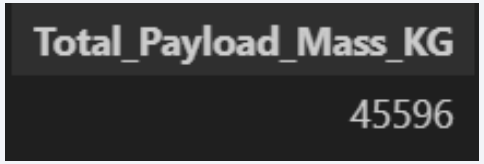
- Task: Display 5 records where launch site begins with the string “CCA”
- Query:
  - %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' LIMIT 5
- Explanation:
  - like ‘CCA%’ selects all records where the launch site starts with CCA.
  - LIMIT 5 displays only the first five records.
- Result:

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

Public Content

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

Public Content

Average Payload Mass by F9 v1.1

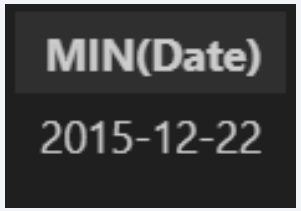
- Task: Display average payload mass carried by booster version F9 v1.1
- Query:
  - %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1’
- Explanation:
  - WHERE clause filters records to display records matching the specified booster version.
  - AVG(PAYLOAD_MASS__KG_)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

AVG(PAYLOAD_MASS__KG_)
2928.4

Public Content

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:



Public Content

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

Public Content

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

Public Content

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

Public Content

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

Public Content

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

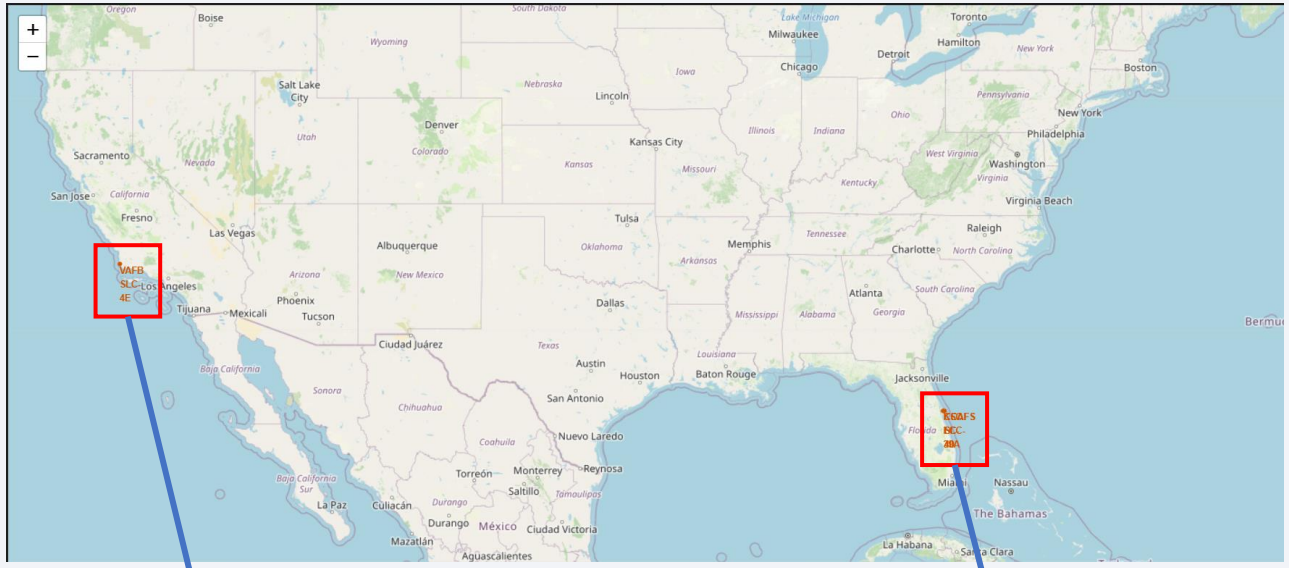
Public Content

Section 3

Launch Sites Proximities Analysis

Public Content

Map of All SpaceX Falcon 9 Launches



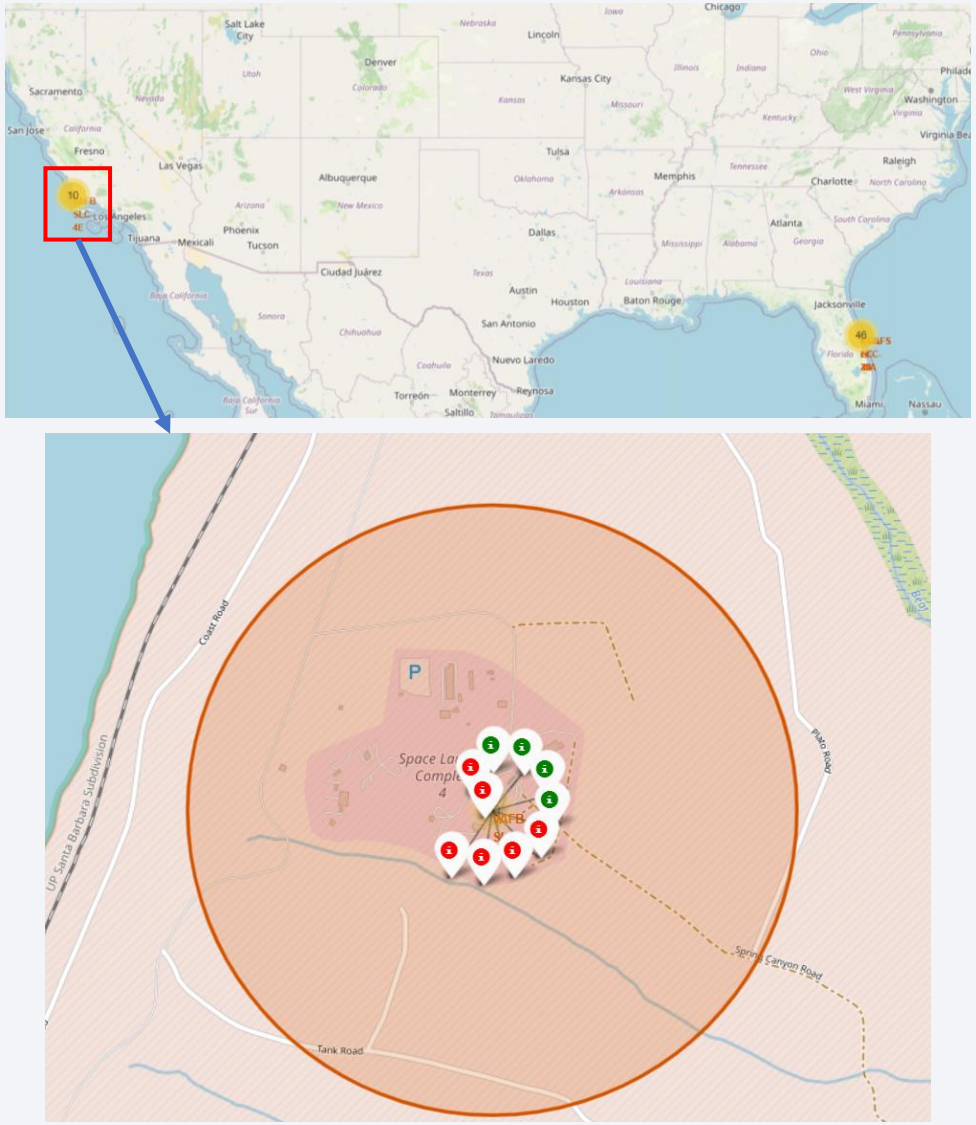
This map shows the location of the four launch sites. The bottom two images are zoomed in to show more detail.

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.

Public Content

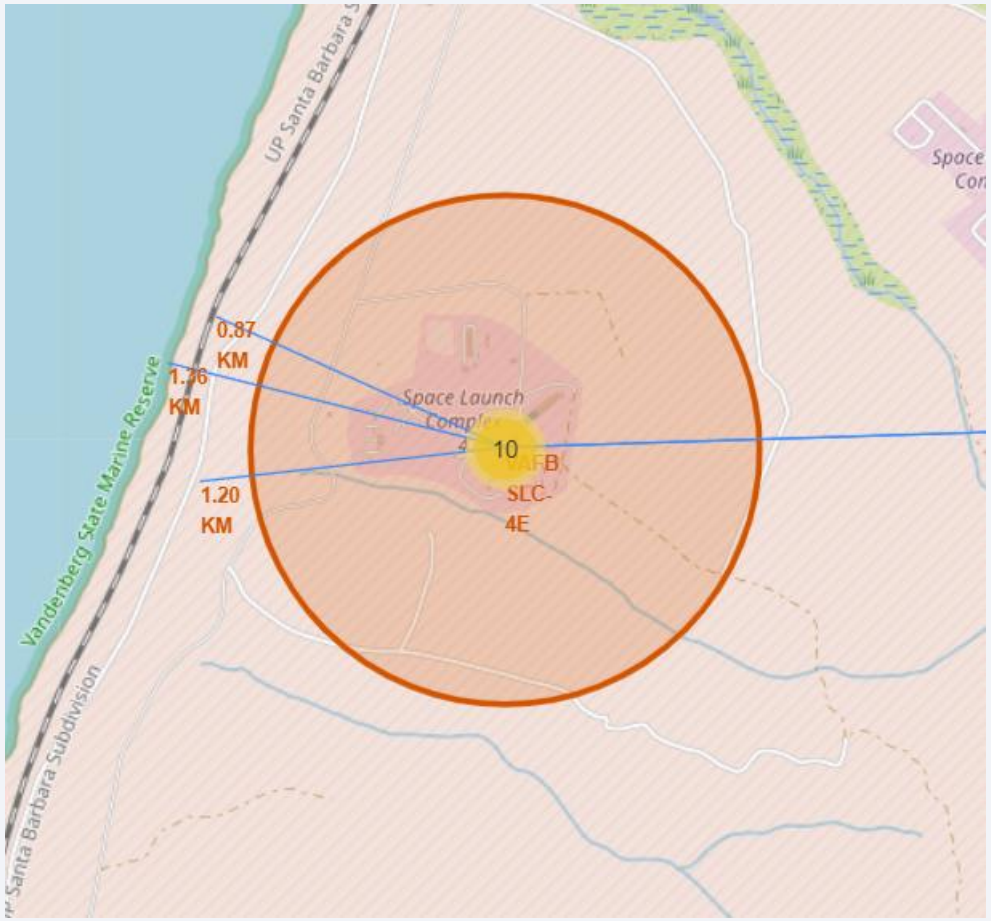
Launch Outcomes By Site



- 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

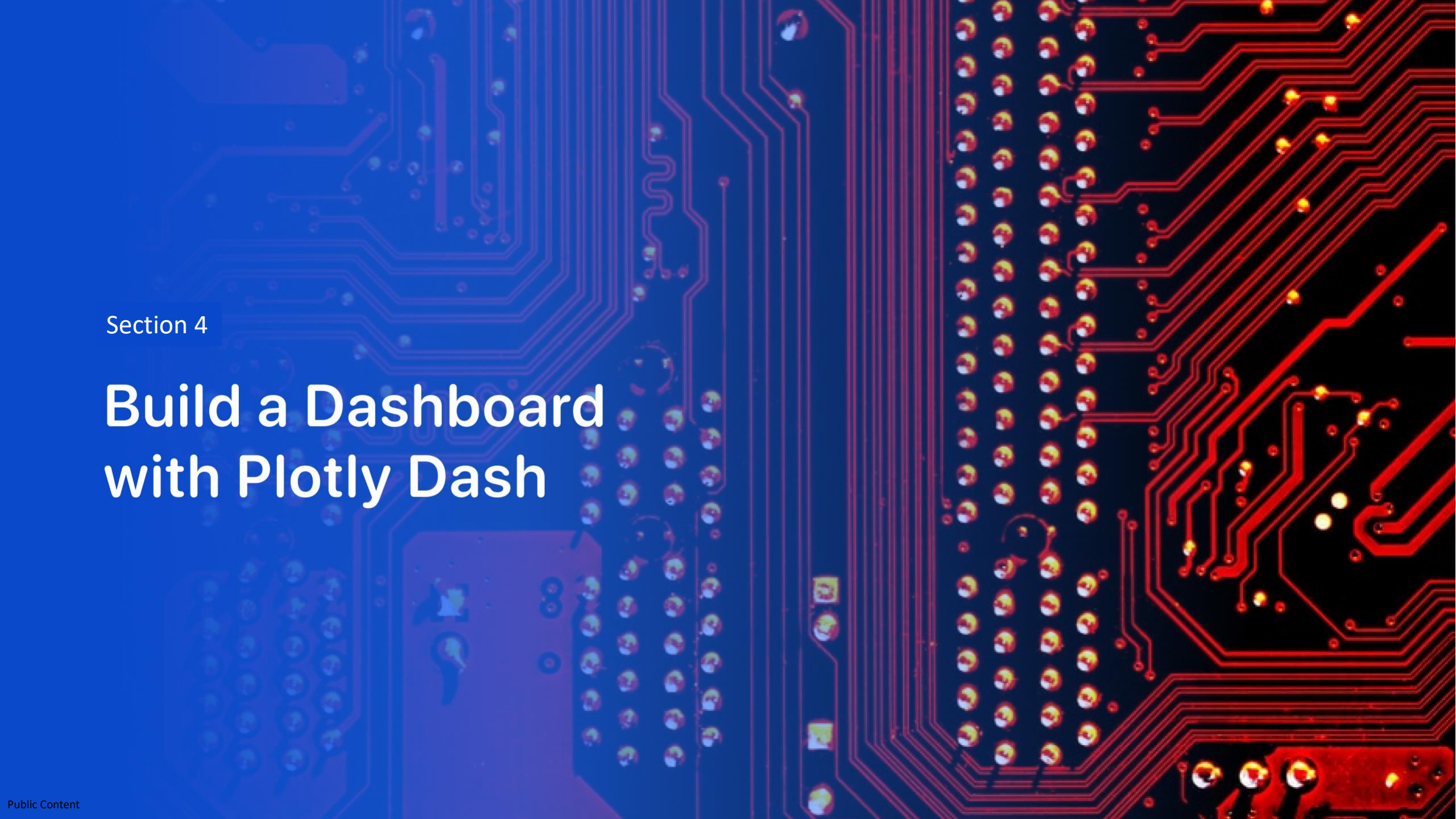
Public Content

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)

Public Content

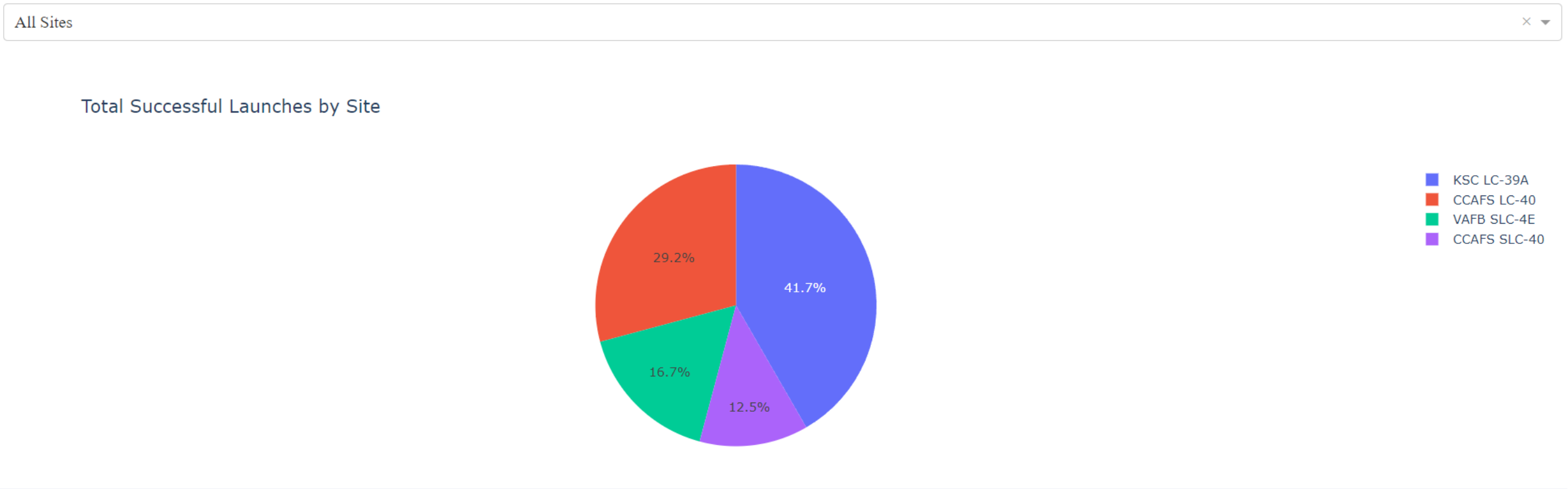


Section 4

Build a Dashboard with Plotly Dash

Public Content

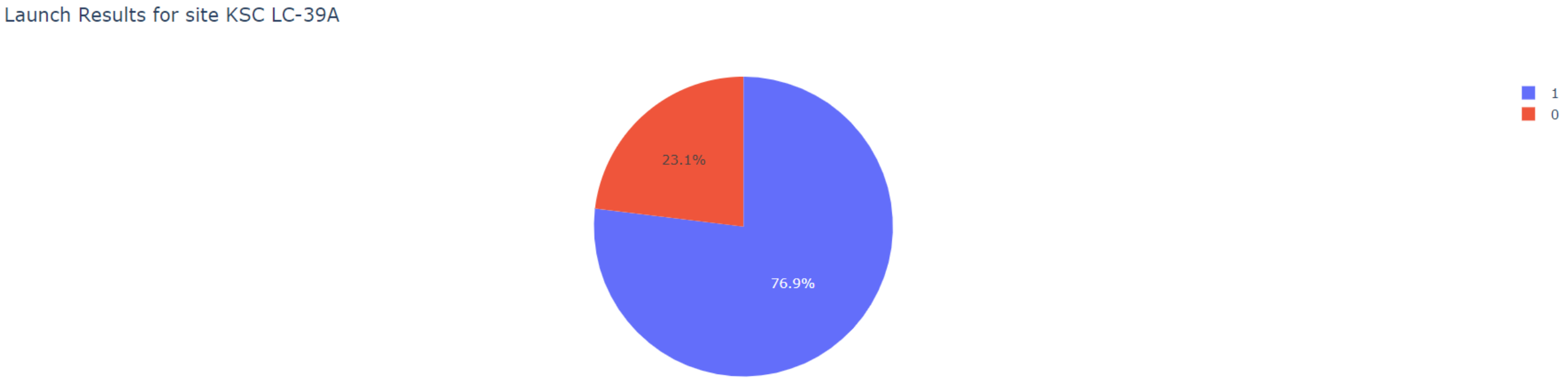
Total Successful Launches, By Site



- 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%

Public Content

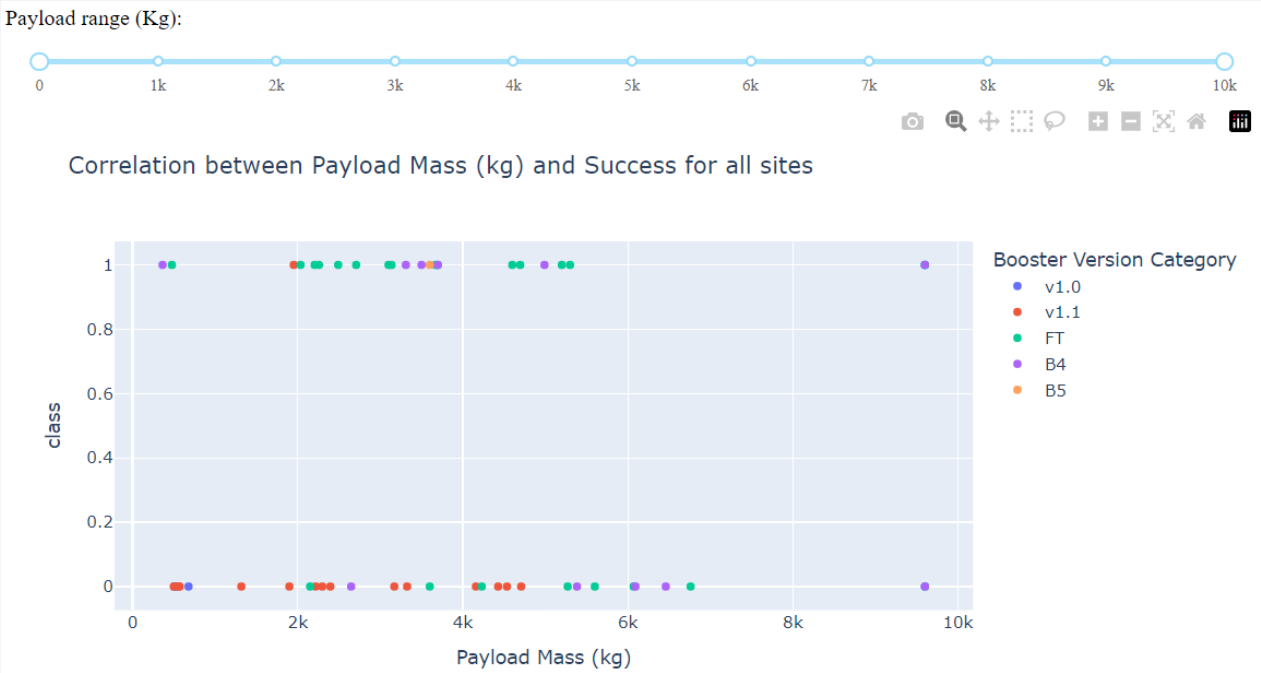
Launch Results for KSC LC-39A

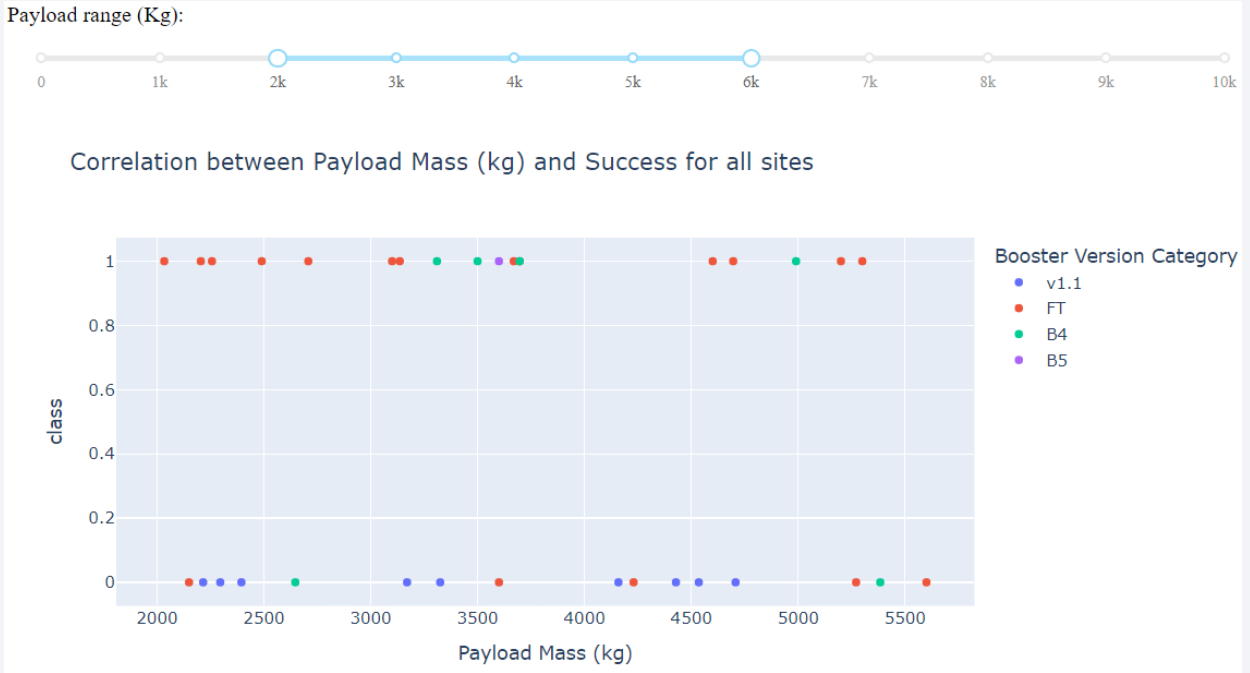


- KSC LC-39A – Launch site with the highest number of successful launches.
- Site has a success rate of 76.9%
- 23.1% of launches at this site failed.

Public Content

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.

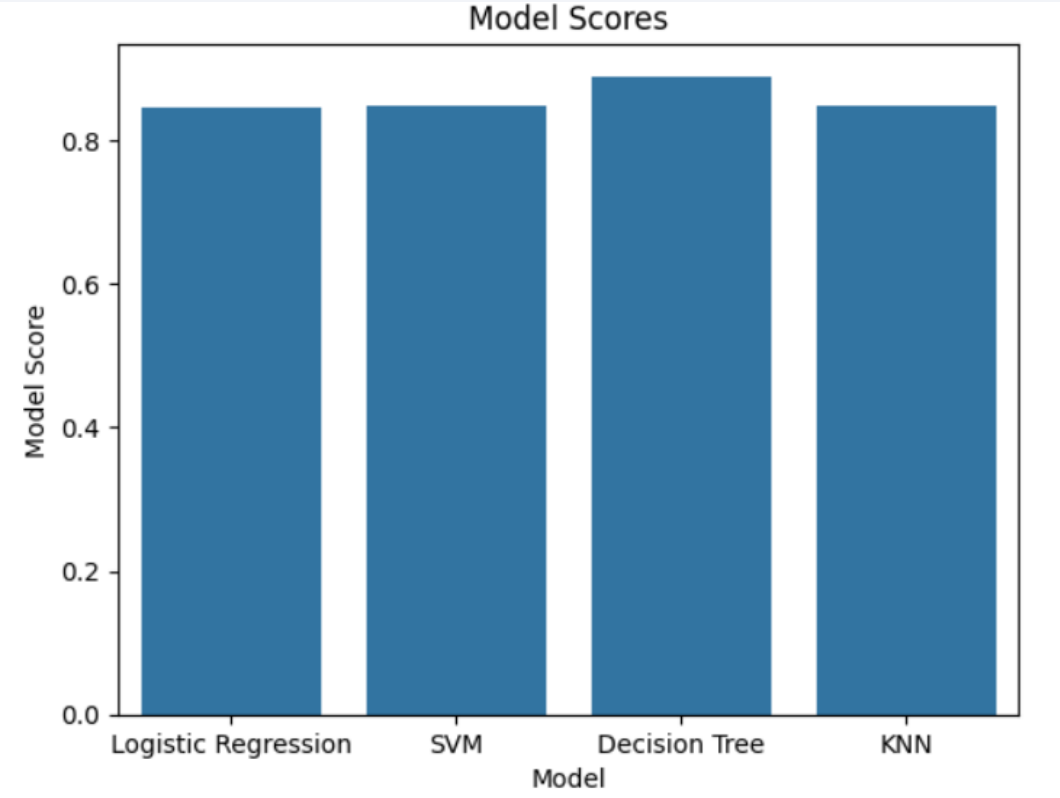
Public Content

Section 5

Predictive Analysis (Classification)

Public Content

Classification Accuracy

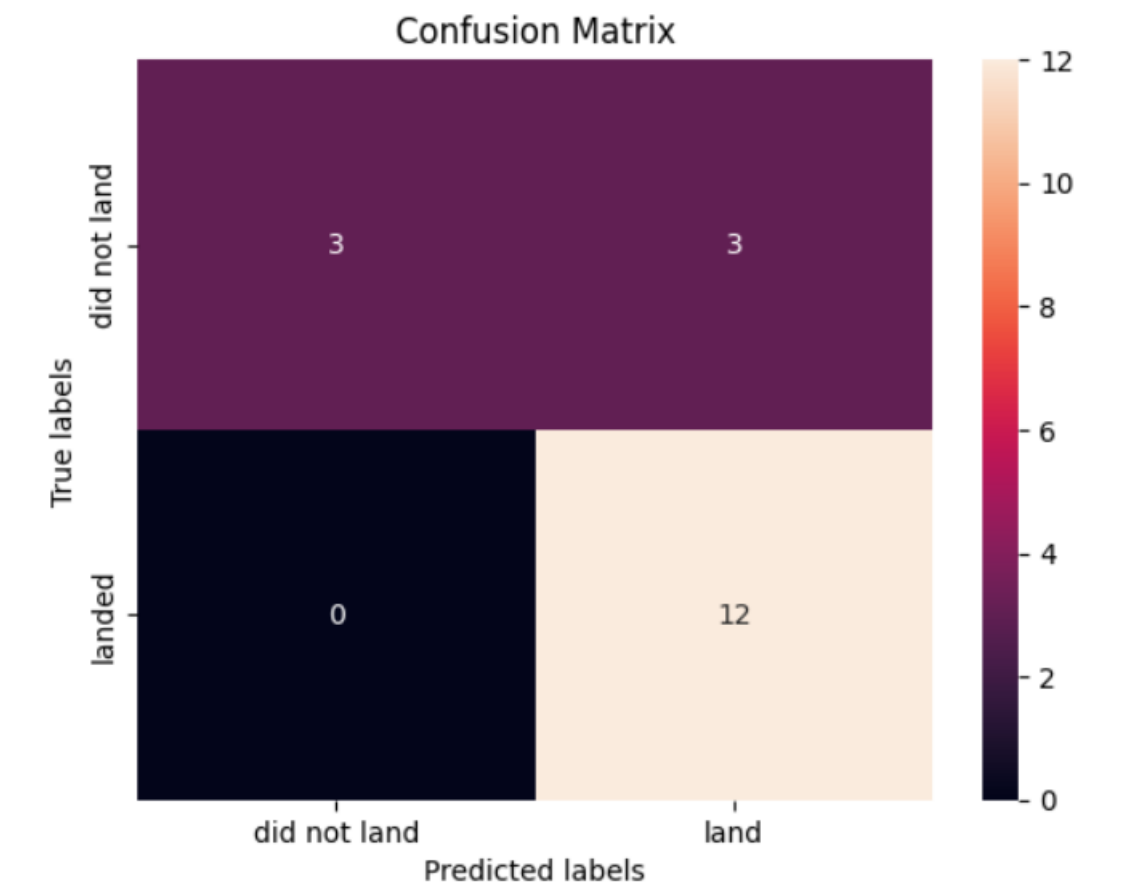


	Model	Model Score	Model Test Data Score
0	Logistic Regression	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Decision Tree	0.889286	0.833333
3	KNN	0.848214	0.833333

- 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.

Public Content

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.

Public Content

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.

Public Content

Thank you!

Public Content