IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

Summary of methodologies

1. Data Collection: Gathering relevant datasets from various sources, including APIs and web scraping techniques, to obtain information pertinent to rocket launches.
2. Data Wrangling: Cleaning and preprocessing the collected data to handle missing values, normalize formats, and prepare the data for analysis.
3. Exploratory Data Analysis (EDA): Analyzing the data to uncover patterns, trends, and relationships using statistical summaries and visualizations.
4. Data Visualization: Creating visual representations of the data using libraries such as Matplotlib, Seaborn, and Plotly to aid in understanding and communicating findings.
5. Model Development: Building and training machine learning models, including Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN), to make predictions based on the data.
6. Model Evaluation: Assessing the performance of the developed models using metrics like accuracy, precision, recall, and confusion matrices to determine their effectiveness.
7. Reporting Results: Compiling the analysis, findings, and model evaluations into a comprehensive report to communicate results to stakeholders.

Summary of all results

- 1. **Exploratory data analysis revealed that launch site and payload mass significantly influence mission success**, with KSC LC-39A showing the highest success rate and certain payload ranges performing more reliably.
- 2. **Interactive visualizations and dashboard components (using Plotly Dash) enabled deeper insights**, allowing users to filter by launch site and payload range, which highlighted trends like higher success rates at specific sites and payloads.
- 3. **Machine learning models—including SVM and Logistic Regression—were able to predict launch success with up to 83% accuracy**, showing that SpaceX launch outcomes can be reasonably predicted using selected features such as site, orbit type, and booster version.
- 4. **The use of Folium for geospatial visualization successfully mapped global launch locations and outcomes**, helping to correlate geographic factors with mission results and giving a clear spatial perspective of SpaceX operations.

Introduction

Background

- The commercial space industry is booming with key players like SpaceX, Virgin Galactic, and Blue Origin.
- SpaceX leads the market by reducing launch costs through reusable first-stage rockets.
- A Falcon 9 launch costs ~$62M vs ~$165M for competitors — primarily due to successful first-stage recovery.

Aims and Objectives

Use public SpaceX data to:

- Analyze launch missions
- Build machine learning models
- Ultimately **predict if the first stage will land successfully**

Section 1

Methodology

Methodology

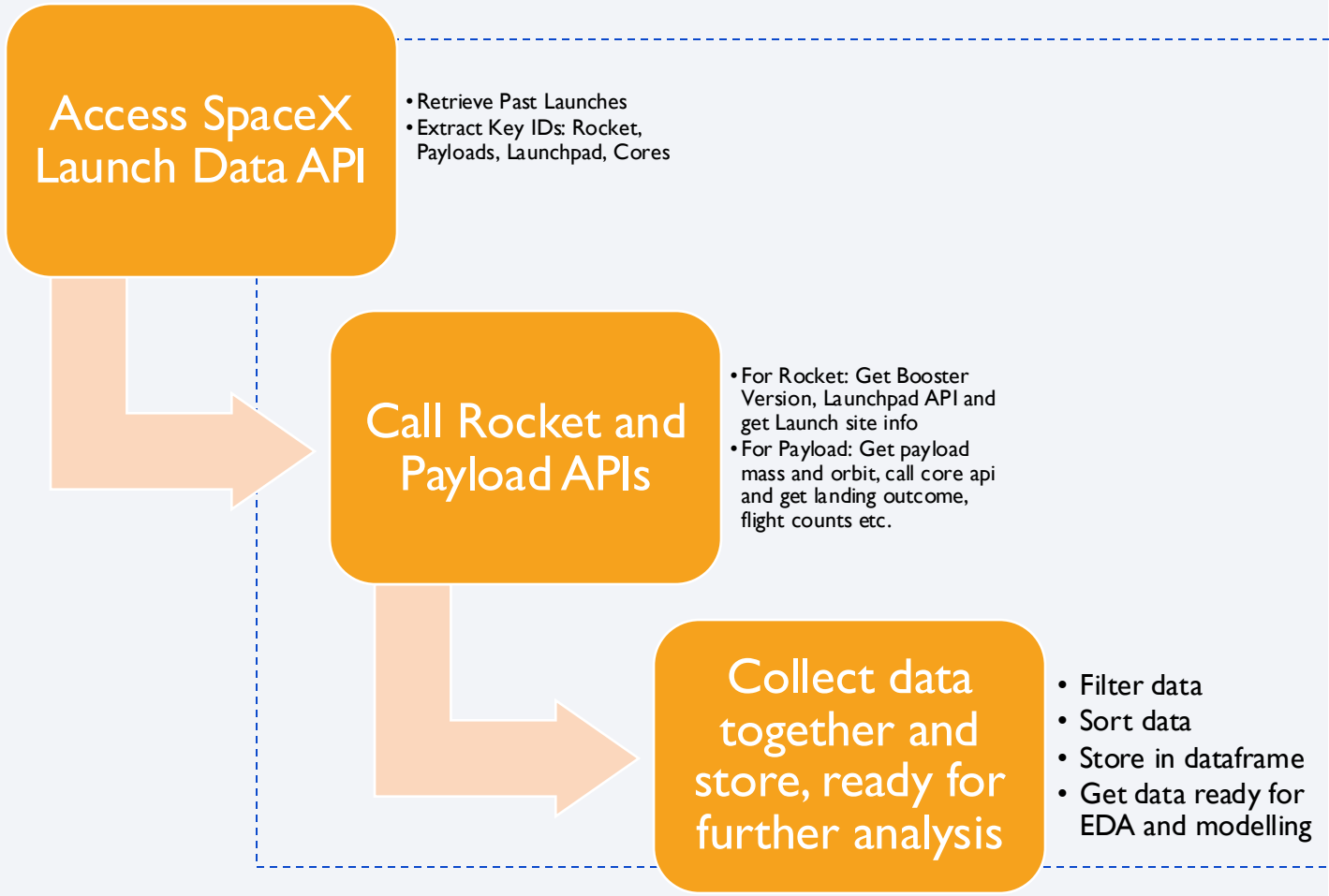
- Data collection methodology
- Perform data wrangling
  - Cleaned inconsistent formats and missing data
  - Normalised features such as payload mass, mission outcomes and orbit types
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Used matplotlib, seaborn and pandas for initial insights
  - Used SQL queries to perform exploration
- Perform interactive visual analytics using Folium to create geospatial maps to show launch sites and Plotly Dash
- Perform predictive analysis using classification models
- Used GridSearchCV for hyperparameter optimization.
- Evaluated using cross-validation, accuracy scores, and confusion matrices

Data Collection

- Describe how data sets were collected.
  - The data was collected using SpaceX REST APIs using tools like BeautifulSoup and the requests library from Python

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
- Source: SpaceX public API
- Data: Launches, rockets, payloads, launchpads, cores
- Method: HTTP GET requests via Python requests
- Extracted IDs → detailed info (booster version, payload mass, orbit, landing outcome)
- Filtered single-core/payload launches; removed Falcon 1
- Converted dates; imputed missing payload mass with mean
- Stored data in lists → combined into pandas DataFrame
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
- https://github.com/SomaanBhatti/CourseraCapstoneProject/blob/main/DataCollection%20pt1.ipynb

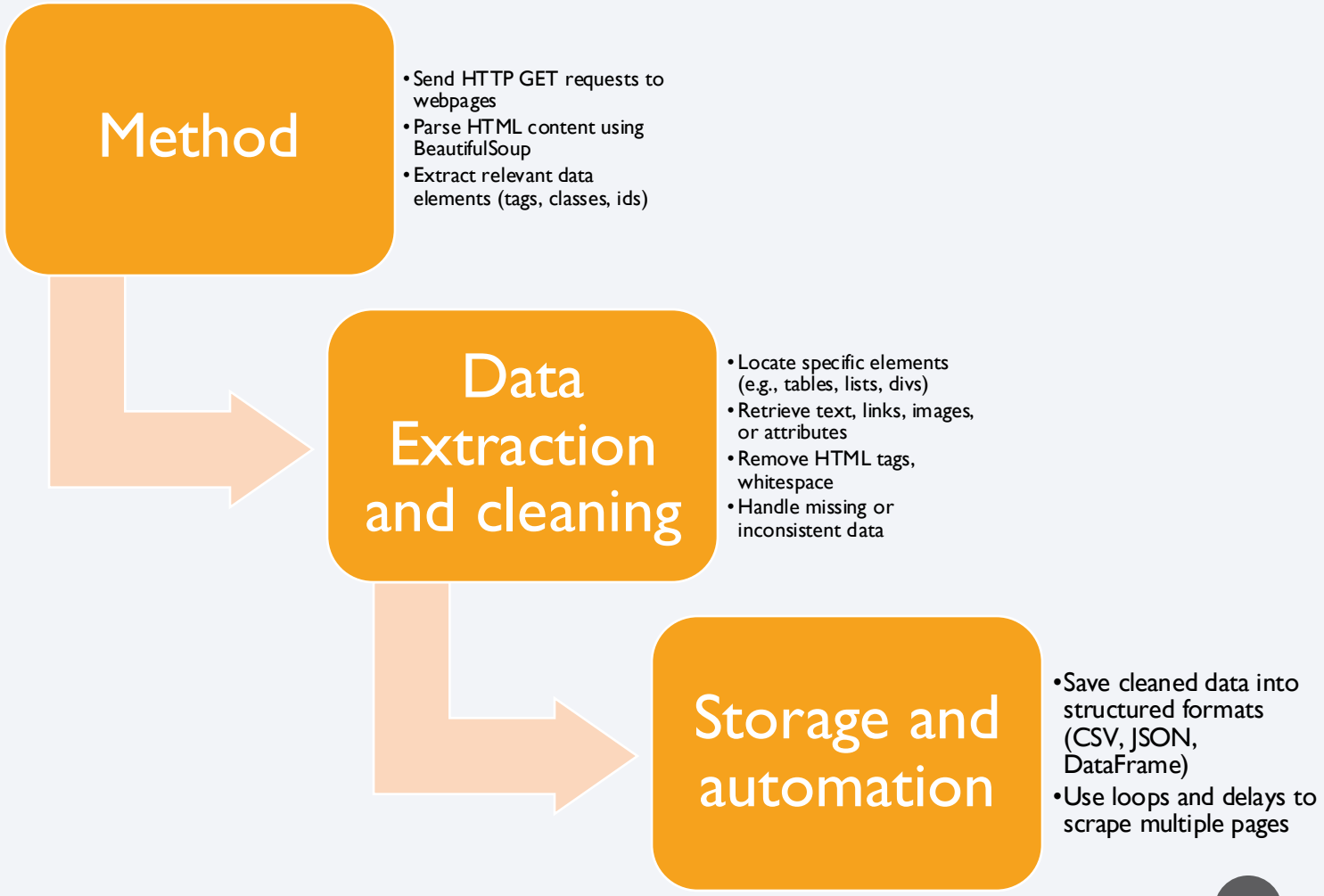


Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

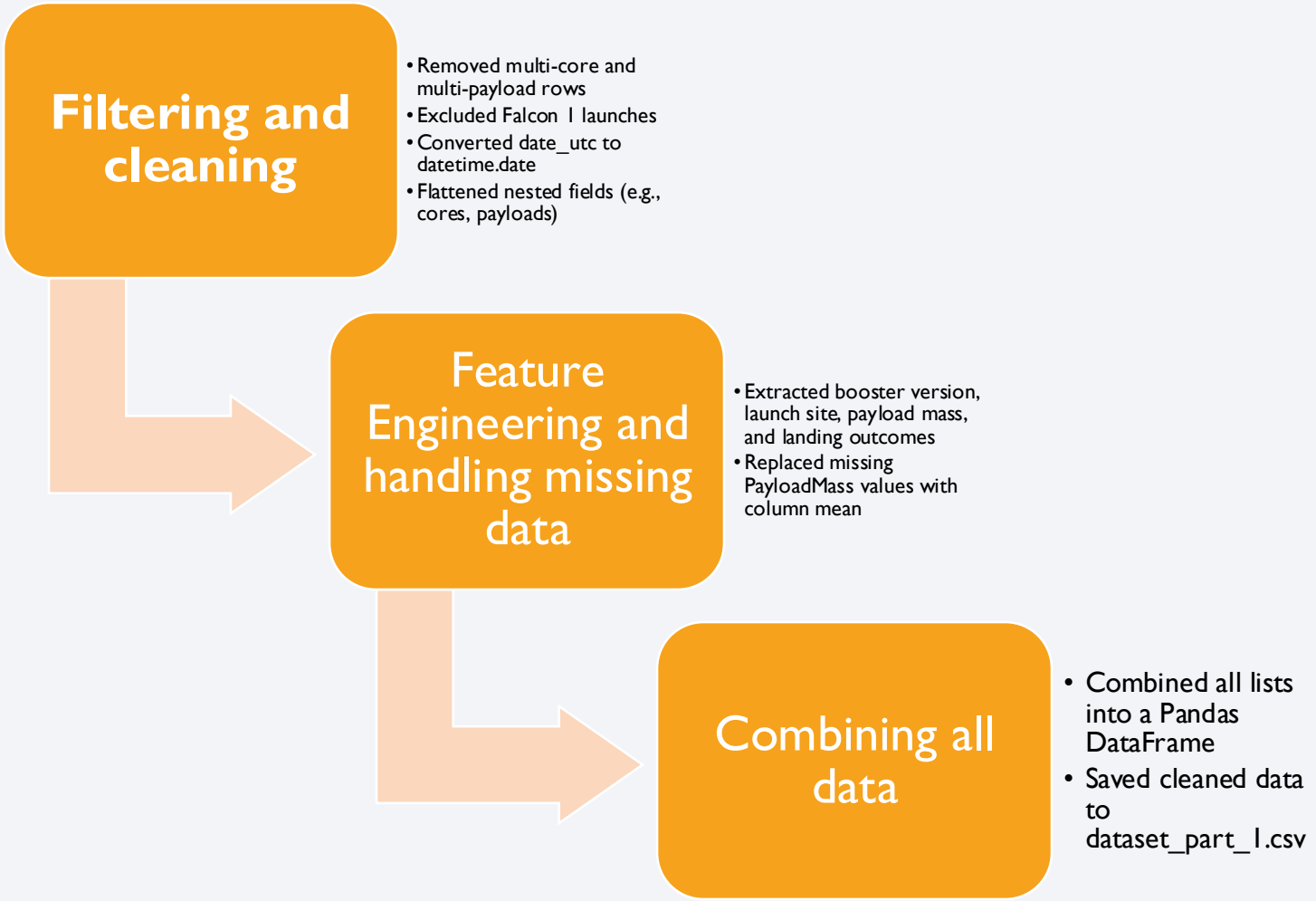
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

- https://github.com/SomaanBhatti/CourseraCapstoneProject/blob/main/webscraping.ipynb



Data Wrangling

- Describe how data were processed You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose
- https://github.com/SomaanBhatti/CourseraCapstoneProject/blob/main/Wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
- https://github.com/SomaanBhatti/CourseraCapstoneProject/blob/main/edadataviz.ipynb

- Data preparation and cleaning involve loading the dataset and ensuring consistency by handling missing values and checking data types
- Exploratory Data Analysis (EDA) uses various visualizations to understand the data better
- Histograms show the distribution of continuous variables, helping identify spread and skewness
- Bar plots display counts or frequencies of categorical variables
- Scatter plots visualize relationships or correlations between two numerical variables, sometimes encoding additional dimensions with color or size
- Box plots compare distributions of numerical variables across categories and help detect outliers
- Python libraries Matplotlib and Seaborn are primarily used for plotting
- Seaborn functions like sns.histplot(), sns.barplot(), sns.scatterplot(), and sns.boxplot() generate these visuals with attractive default styles
- Plots are customized with titles, axis labels, legends, colors, plot sizes, and palettes for clarity and aesthetics
- Visualizations are interpreted to extract insights such as trends, outliers, and relationships, guiding further analysis or modeling steps

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose
- https://github.com/SomaanBhatti/CourseraCapstoneProject/blob/main/eda%20with%20sql.ipynb

- Data is queried from a PostgreSQL database using the psycopg2 Python library connection
- Basic SELECT statements retrieve data from tables for exploration
- Filtering with WHERE clauses narrows down results based on conditions like specific age ranges or categories
- GROUP BY is used to aggregate data, such as counting records per group or calculating averages
- ORDER BY sorts the query results by one or more columns in ascending or descending order
- JOIN operations combine data from multiple tables to enrich the dataset and enable more complex analysis
- Aggregate functions like COUNT(), AVG(), MAX(), and MIN() summarize data within groups
- LIMIT restricts the number of rows returned for quick preview of data samples
- Subqueries are used to perform nested queries to filter or calculate intermediate results before the main query
- Queries include date/time filtering to select records within specific timeframes or periods
- Results of queries are loaded into pandas DataFrames for further analysis and visualization within Python
- These SQL queries help in exploratory data analysis by extracting relevant subsets and aggregated views of the data to understand distributions and relationships

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
- https://github.com/SomaanBhatti/CourseraCapstoneProject/blob/main/visual%20analytics%20folium.ipynb

- Created a base Folium map centered on specific geographic coordinates with an appropriate zoom level
- Added Markers to the map to pinpoint exact locations, often with pop-up labels to provide additional information about the point
- Used Circle Markers to represent locations with customizable radius and color, highlighting importance or magnitude of a feature
- Added Circles with specific radius sizes on the map to visualize areas around certain points, helping show spatial influence zones
- Plotted Polylines (lines) connecting multiple geographic points to represent paths or routes between locations
- Utilized Layer Controls to toggle different map layers or groups of markers on and off for interactive exploration
- Added Popups and Tooltips to markers and circles for displaying contextual information when users click or hover over map objects
- Customized map objects with different colors, fill opacity, and radius to enhance visual differentiation and clarity on the map
- Used Feature Groups to organize related markers and shapes into layers for better map management and interactivity
- The map objects collectively provide a rich, interactive spatial visualization of the data, enabling geographic pattern recognition and exploration

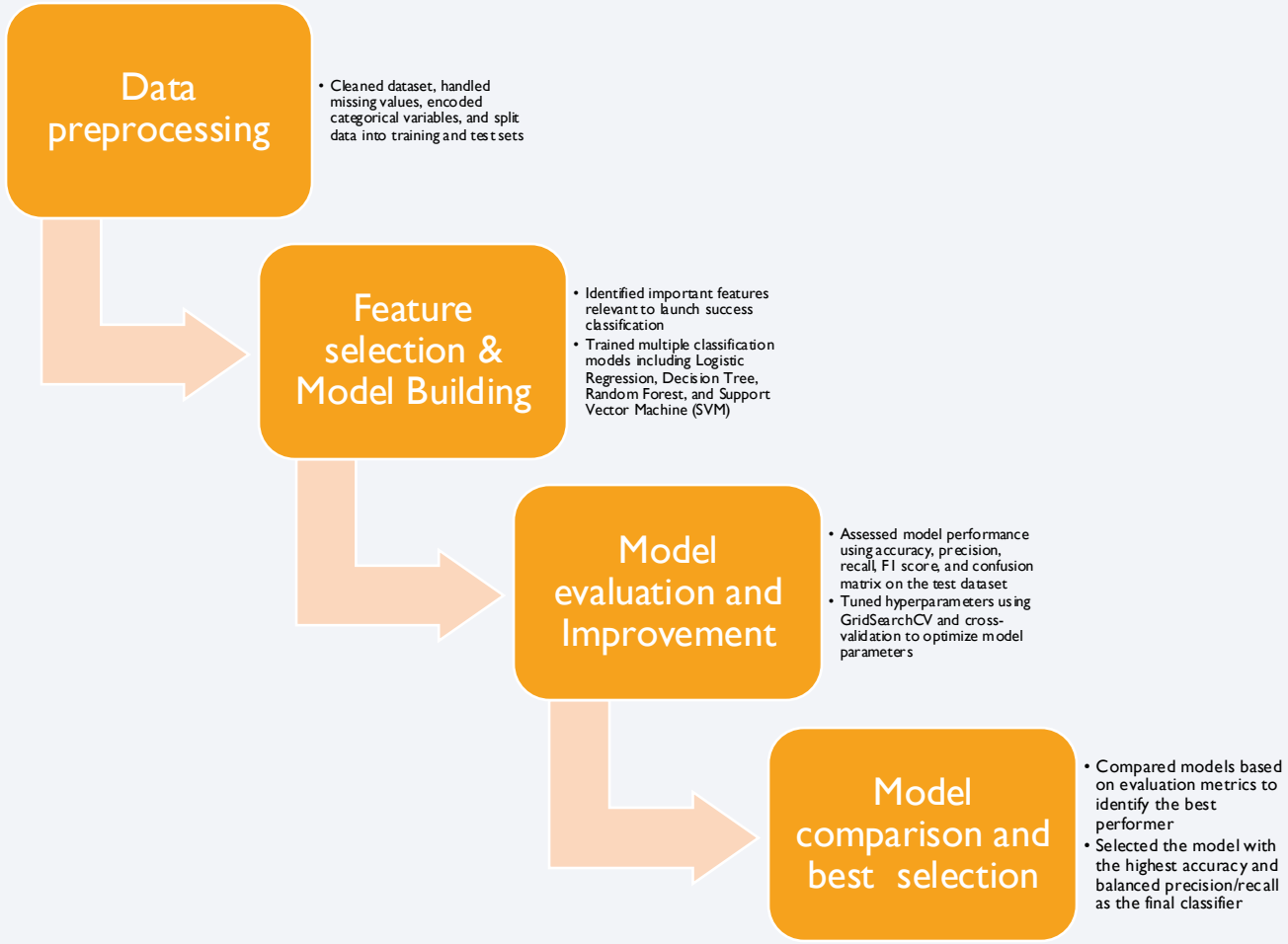
Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- https://github.com/SomaanBhatti/CourseraCapstoneProject/blob/main/spacex-dash-app.py

- Added a dropdown input component to select the launch site, enabling filtering of data based on the selected site
- Created a pie chart that dynamically updates to show the count or proportion of successful launches for the selected launch site from the dropdown
- Implemented a callback function that listens to the dropdown selection and updates the success pie chart accordingly
- Added a range slider input to select a payload mass range, allowing users to filter data by payload weight
- Built a scatter plot chart displaying the relationship between payload mass and launch success, with points colored or marked by booster version or other categories
- Implemented a callback function that listens to both the dropdown site selection and the payload range slider input to update the scatter plot dynamically
- Included interactive features such as hover tooltips on the scatter plot to provide detailed information about individual launches
- The dashboard integrates these input components and visualizations to allow users to explore launch success rates by site, payload range, and booster version
- This interactive dashboard enables answering key questions about launch performance and success factors through real-time visual updates based on user input

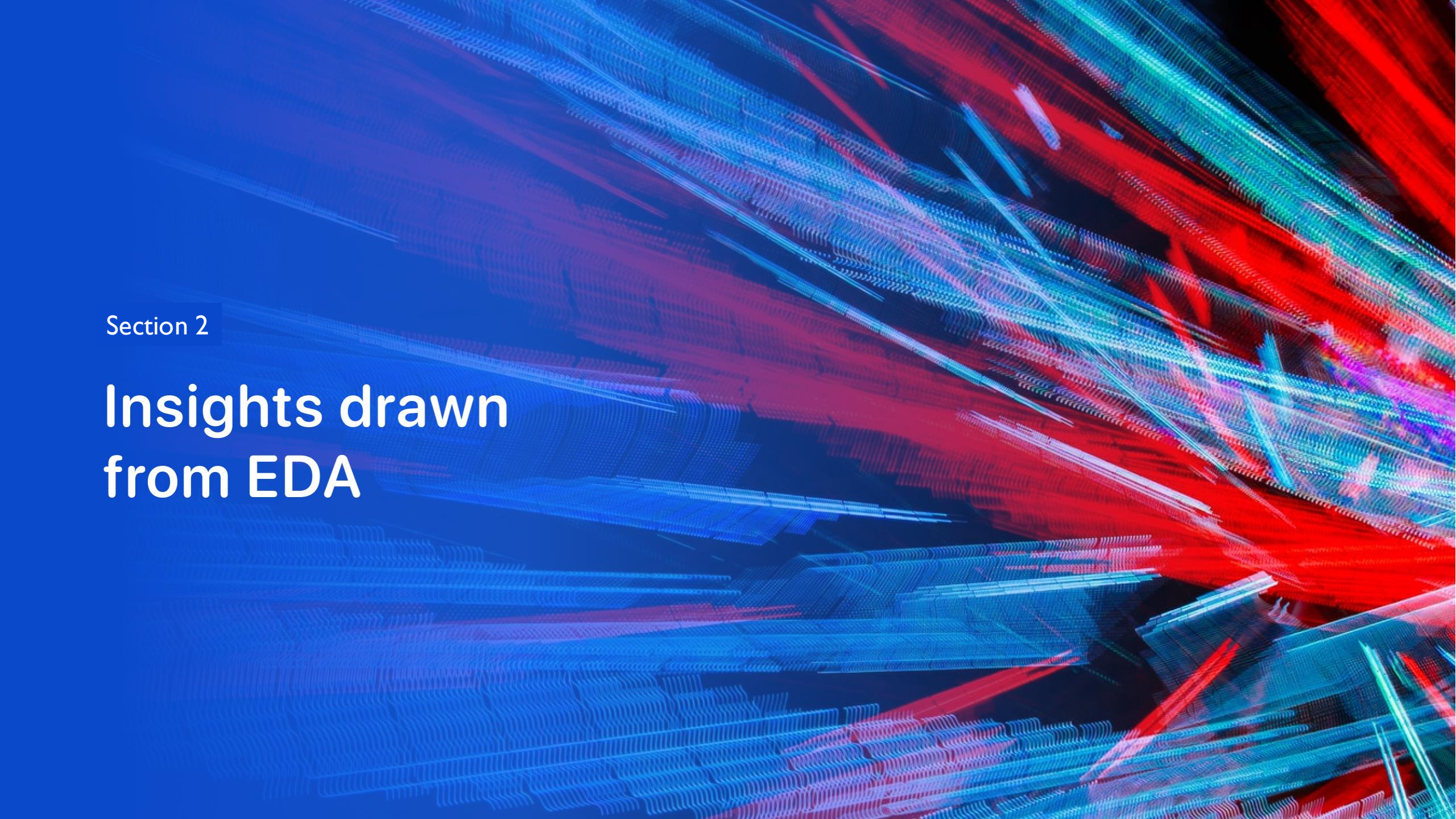
Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- https://github.com/SomaanBhatti/CourseraCapstoneProject/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



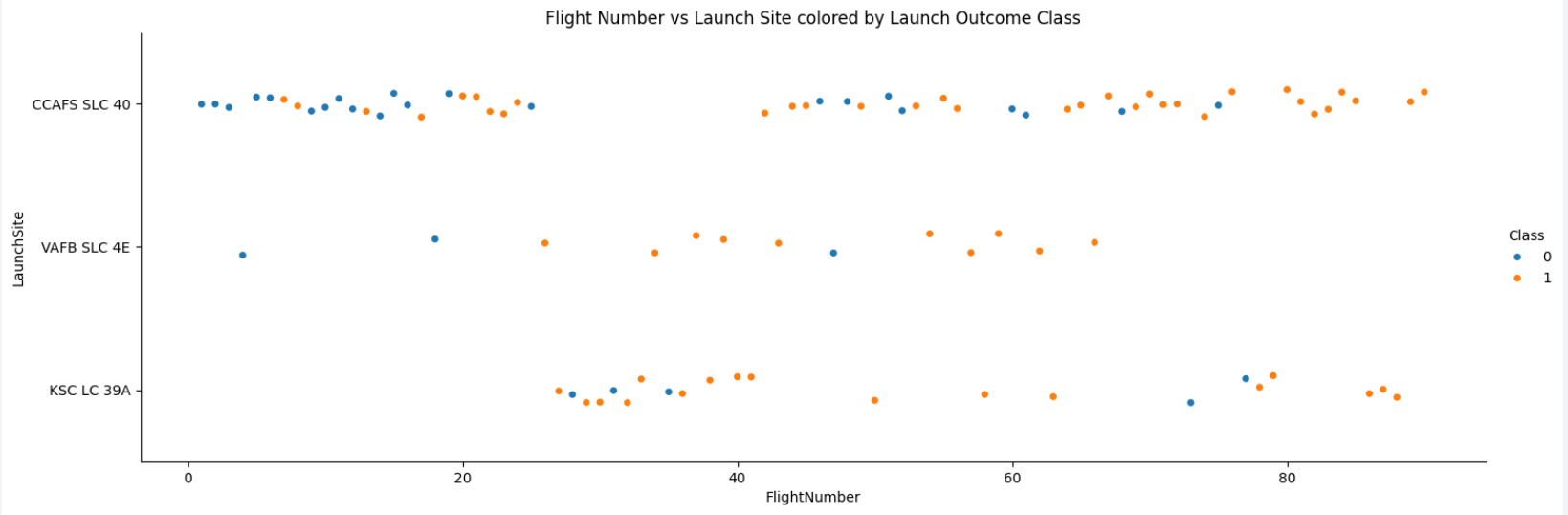
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

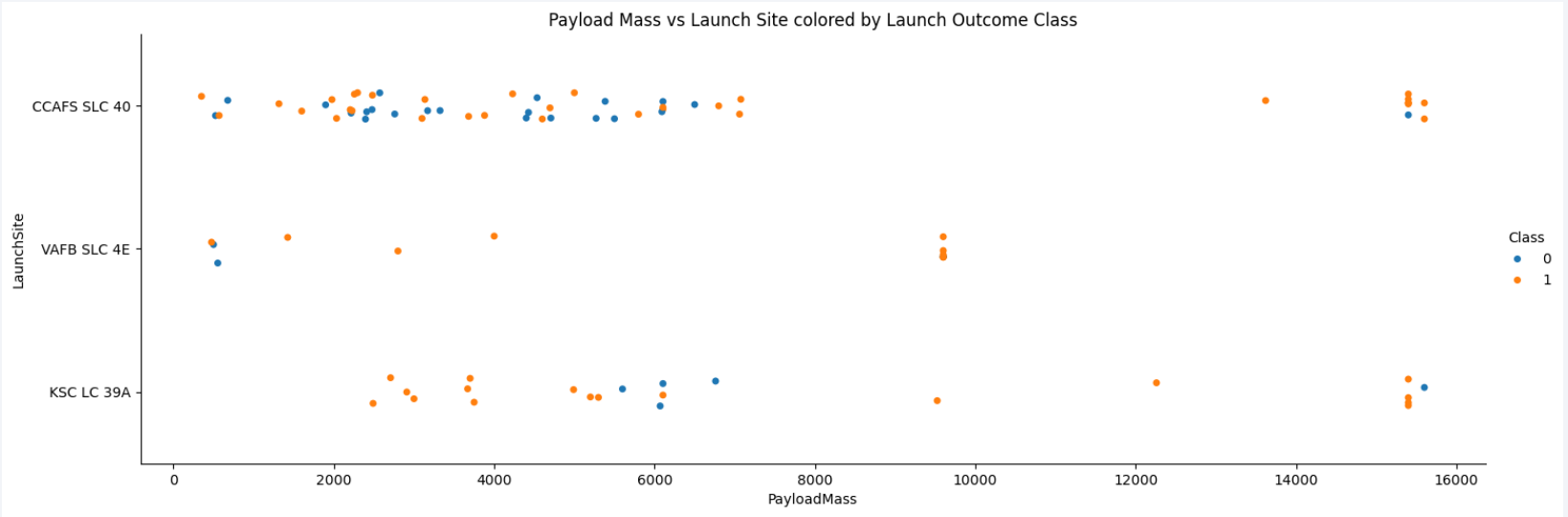
- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations



The launch site represents different geographical locations where SpaceX conducts its launches.
The flight number typically indicates the chronological order or count of total flights conducted by SpaceX over time.
When analyzing the relationship between launch site and flight number, different launch sites have varying numbers of flights associated with them, reflecting the frequency of launches at each site.
Launch sites with higher flight numbers generally indicate more established or frequently used locations.
Plotting flight number against launch site can reveal which sites have hosted more flights and how flight activity has progressed over time at each site.
This relationship helps to understand the operational scale and preference for launch sites, as well as temporal trends in launch activity distributed by location.

Payload vs. Launch Site

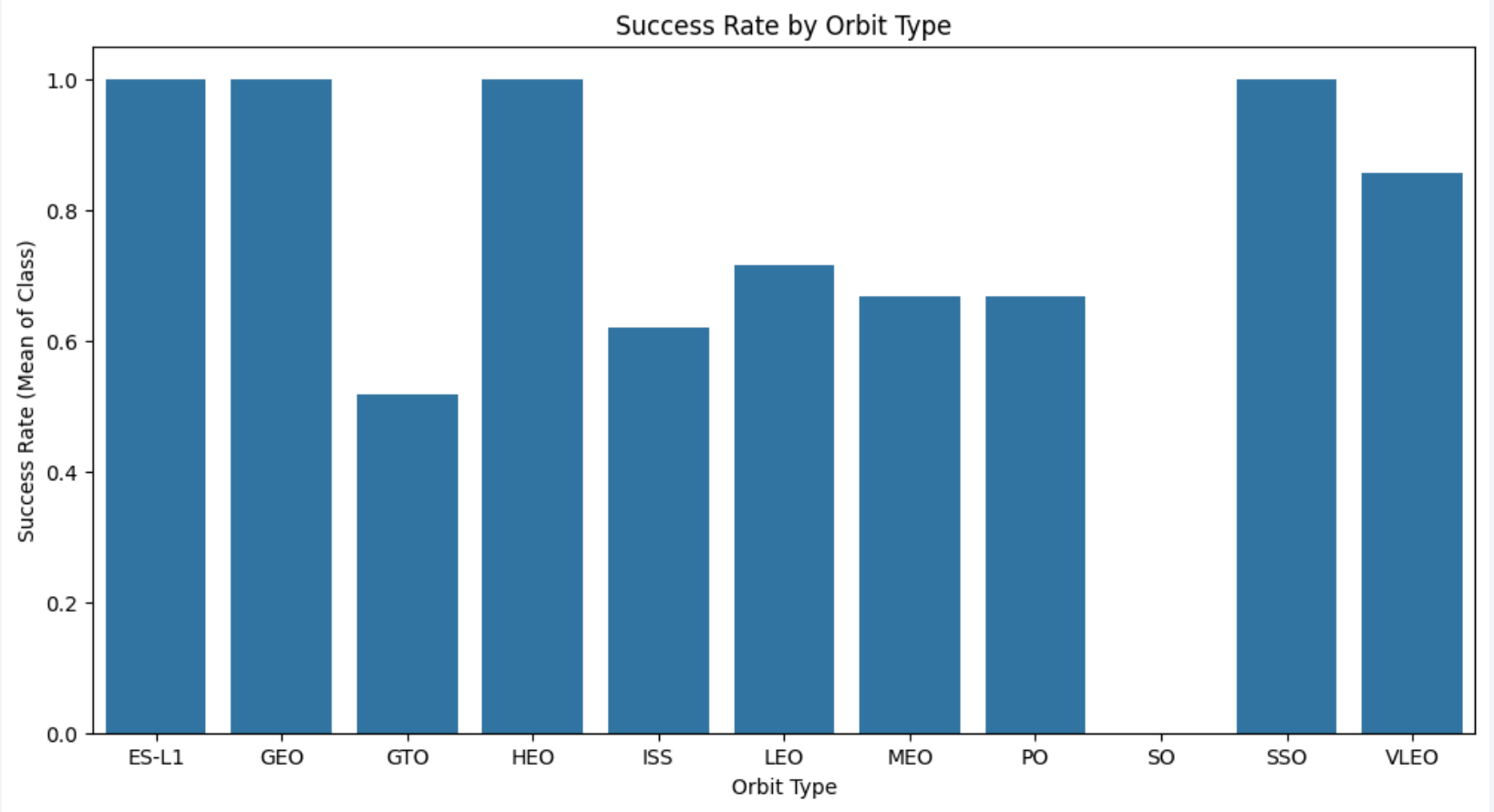
- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations



The payload represents the weight or mass of the cargo being launched by SpaceX. The launch site indicates the specific geographical location where the launch takes place.
When analysing the relationship between payload and launch site, different launch sites handle varying ranges of payload weights, reflecting their operational capabilities and mission profiles.
Some sites may frequently launch heavier payloads due to infrastructure or mission type, while others may focus on lighter payloads.
Visualizing payload distribution across launch sites helps identify which locations support heavier or lighter payload missions.
This comparison provides insight into how launch infrastructure and site selection relate to the types of payloads being launched.

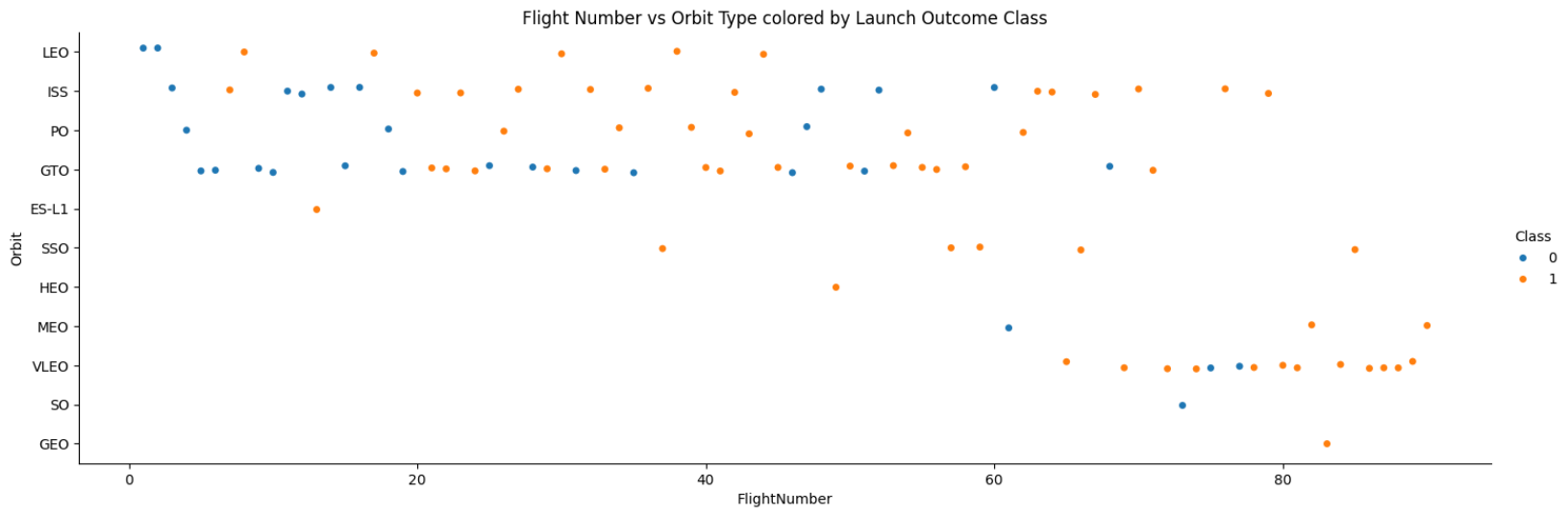
Success Rate vs. Orbit Type

- The orbit type refers to the destination orbit the payload is intended to reach, such as LEO (Low Earth Orbit), GTO (Geostationary Transfer Orbit), or others.
- The bar chart of success rate by orbit type shows how frequently launches to different orbits have succeeded versus failed. Each bar represents an orbit category and its associated success rate, calculated as the proportion of successful launches out of total attempts for that orbit.
- This visualization highlights which orbit types have higher or lower reliability based on past performance.
- Some orbits may have higher success rates due to simpler trajectories or more routine missions, while others may be more complex and risk-prone.
- The chart helps identify patterns in mission reliability across different orbit destinations, offering valuable insight into SpaceX's operational strengths and risks associated with each orbit type.



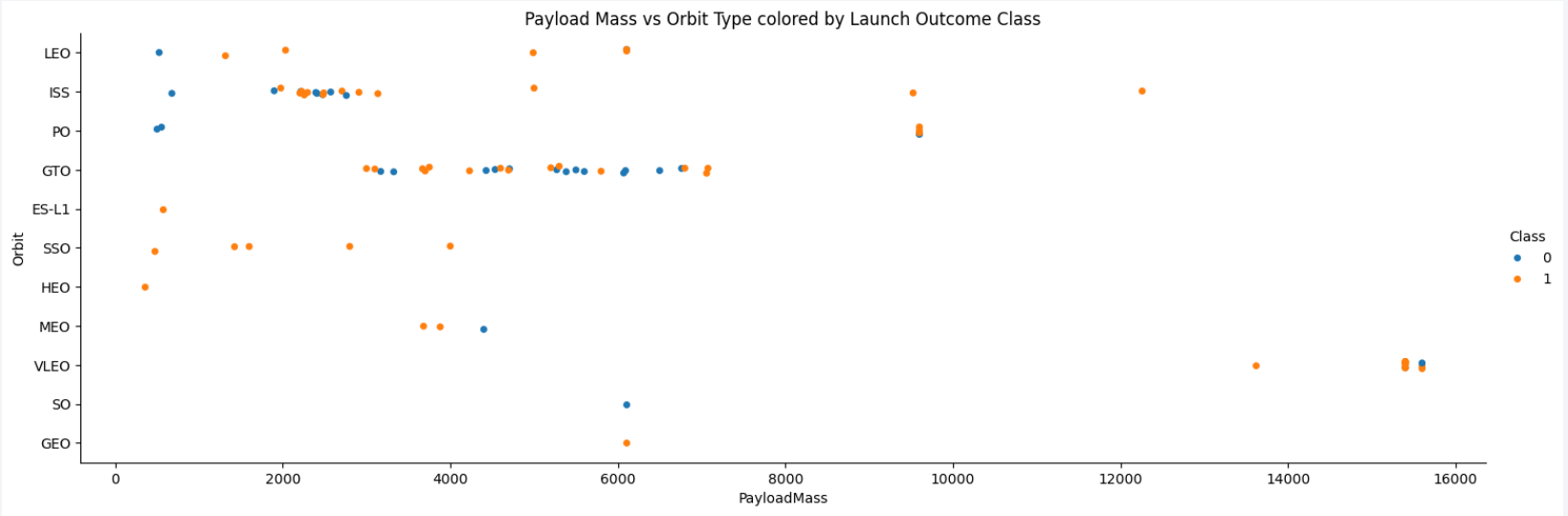
Flight Number vs. Orbit Type

- The flight number indicates the sequential order of SpaceX launches over time. The orbit type represents the target orbit for the payload, such as LEO, GTO, or MEO.
- Plotting flight number against orbit type reveals how the variety or distribution of mission types has changed over time. It shows which orbit types were targeted in earlier missions compared to later ones.
- This relationship can highlight trends, such as whether SpaceX has increased launches to higher or more complex orbits as the number of flights has grown.
- It also helps identify when new orbit types were introduced into the launch portfolio.
- The plot provides insight into the evolution of mission complexity and diversification as SpaceX has matured operationally.



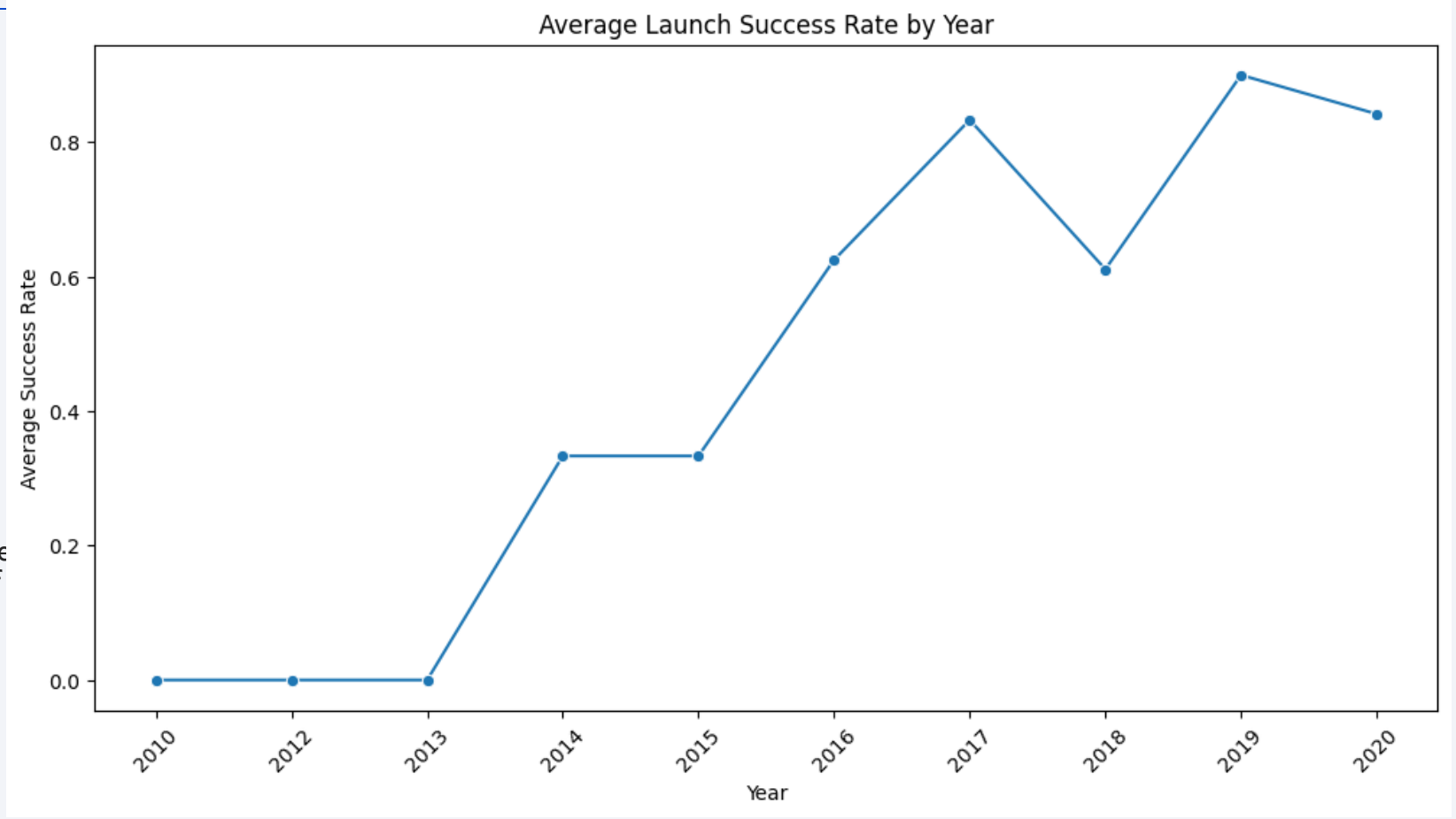
Payload vs. Orbit Type

- Analysing payload vs orbit type helps reveal how payload mass varies depending on the orbital destination.
- Different orbit types typically support different payload capacities due to varying energy and altitude requirements. For example, GTO missions often carry lighter payloads compared to LEO missions, due to the higher energy needed to reach geostationary transfer orbit.
- The plot shows the distribution of payload masses across orbit types and helps identify which orbits are associated with heavier or lighter cargo.
- This relationship provides insight into how mission goals and orbital mechanics influence payload planning and constraints.



Launch Success Yearly Trend

- The launch success rate yearly shows how the proportion of successful launches has changed over time, grouped by year.
- Each data point represents the percentage of successful launches out of the total number attempted in a given year.
- This trend helps visualize SpaceX's reliability improvements or fluctuations in performance over the years.
- A rising trend in success rate suggests growing operational experience, better technology, and more efficient processes. Any dips may indicate periods of testing, experimentation, or anomalies.
- The yearly success rate is a key metric to assess SpaceX's progress and consistency in achieving successful missions over time.



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

First_Successful_Ground_Pad_Landing
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here
- 1 failure in-flight
- 100 Successes

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- Present your query result with a short explanation here

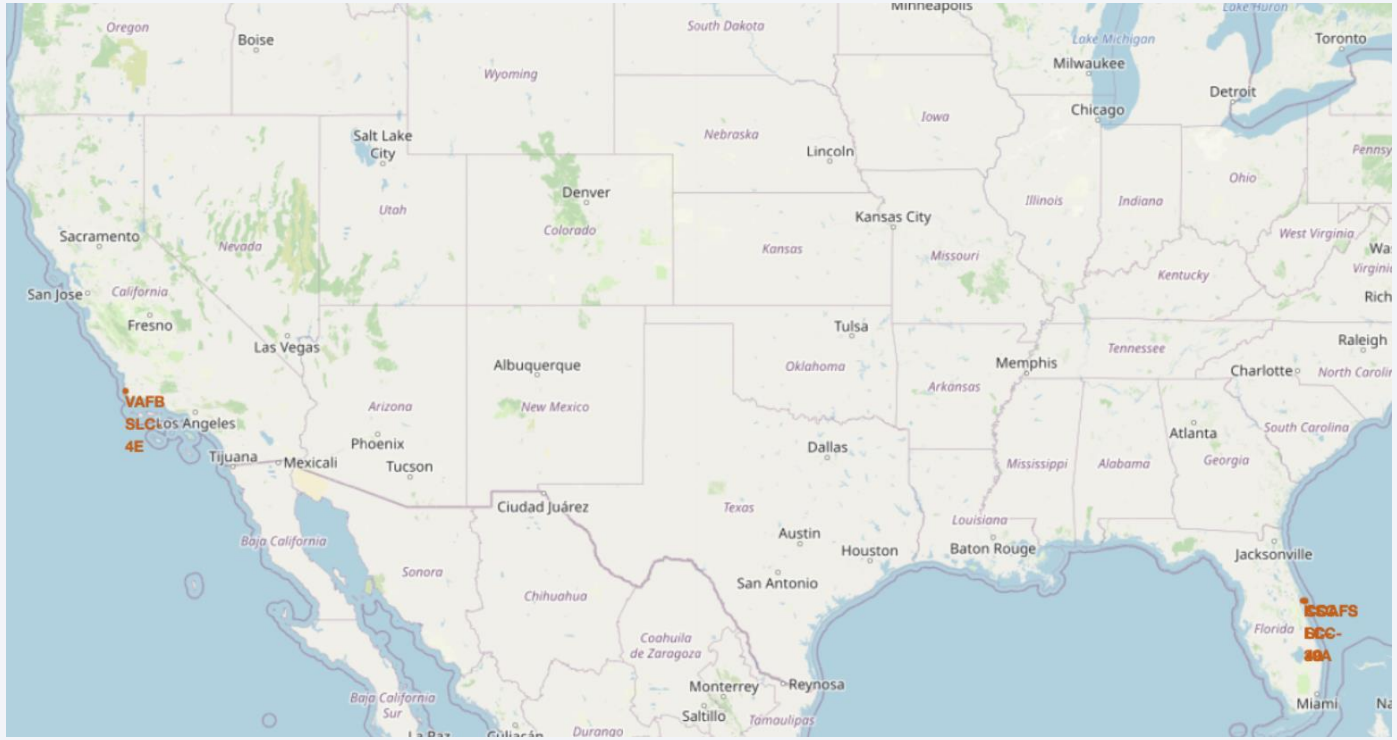
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

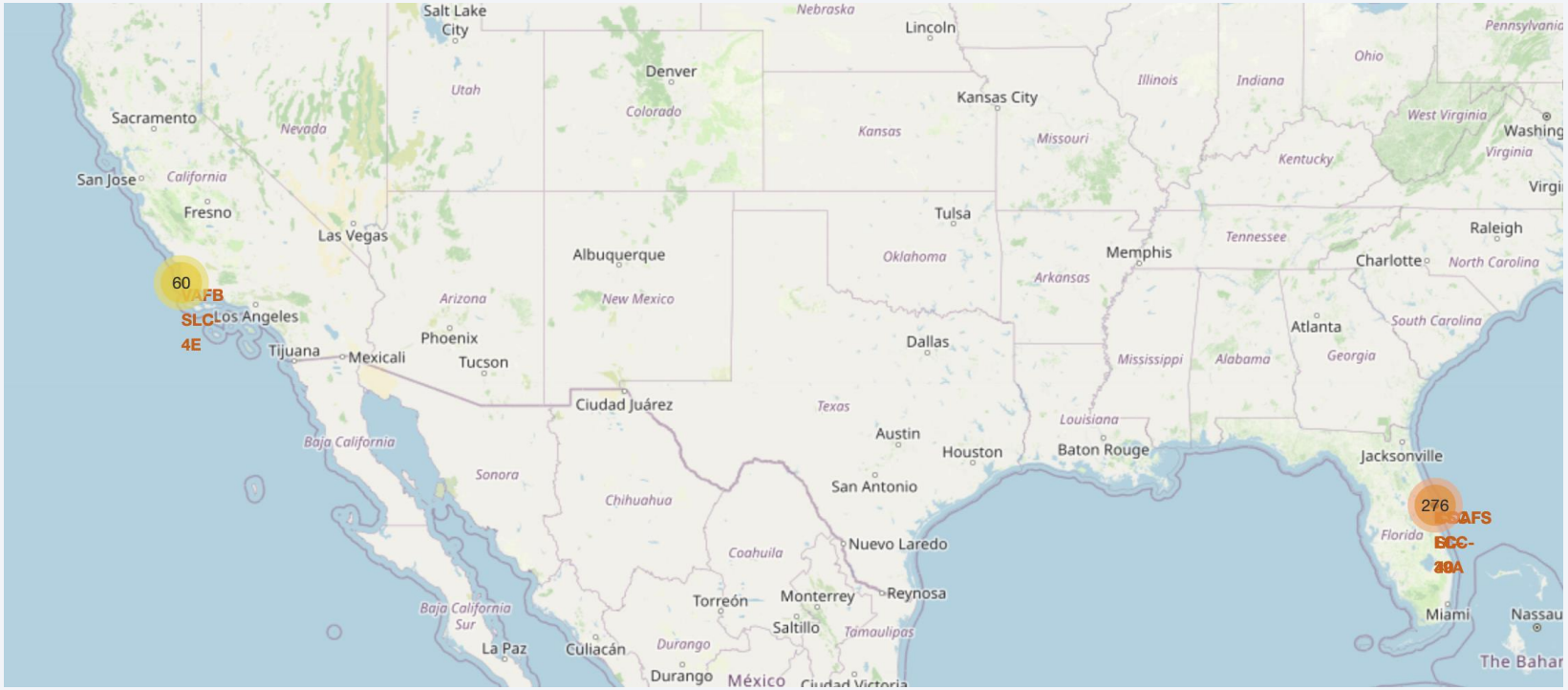
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot



- Each **marker** represents a SpaceX launch site, such as CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A, and others.
- The launch sites are distributed across different parts of the United States, with some located on the east coast (like Cape Canaveral), some on the west coast (like Vandenberg), and others near the Gulf of Mexico or central Florida.
- This distribution supports a range of orbital trajectories, from polar to geostationary.

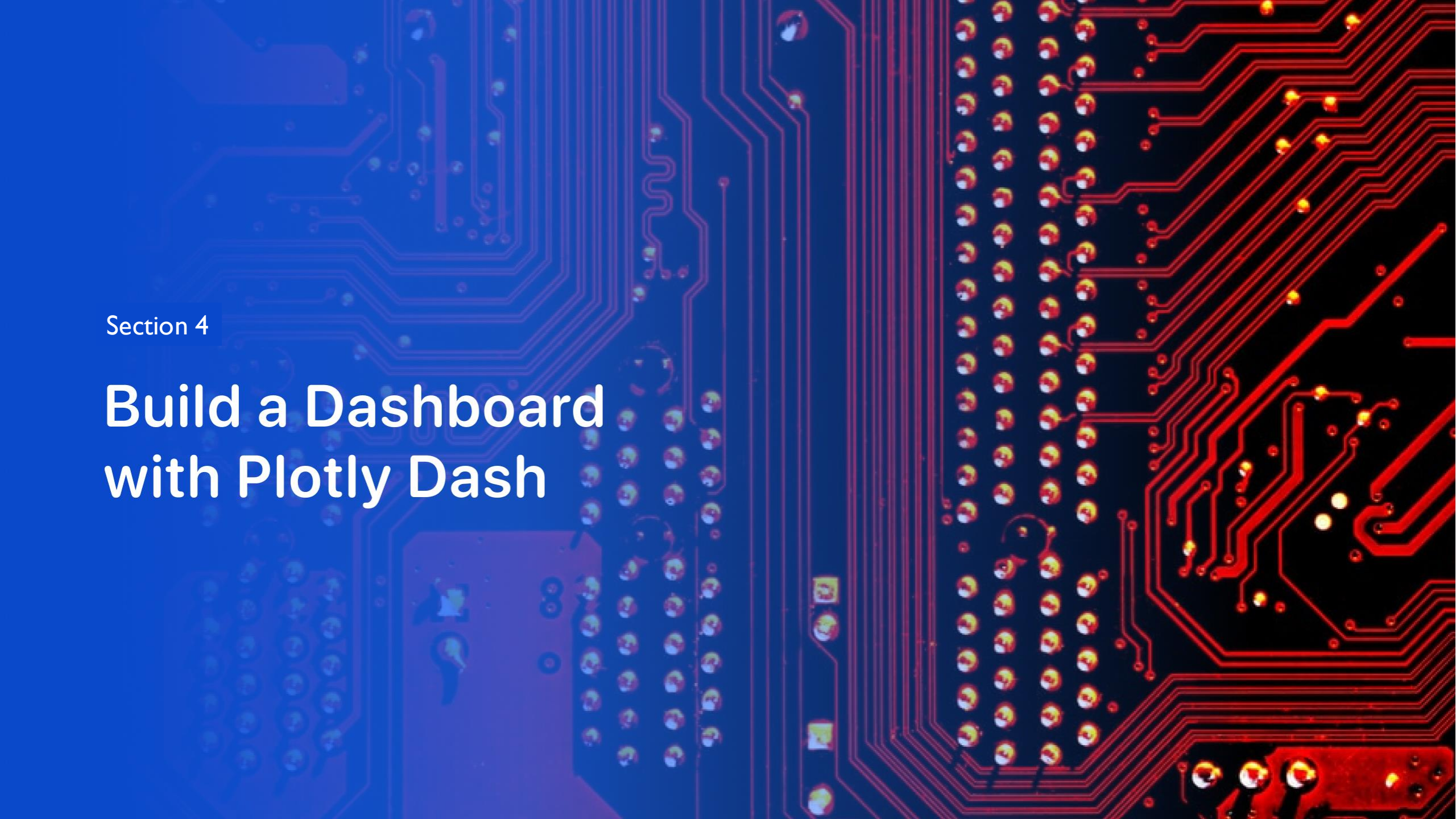
Colour Labeled launch outcomes

- Sites with a high concentration of **green markers** indicate a strong history of successful launches.
- Sites that show **mixed colors** suggest variability in performance or earlier developmental/testing phases.
- If a site has more **red markers**, it may highlight areas of concern, technical challenges, or learning curves during early launches.

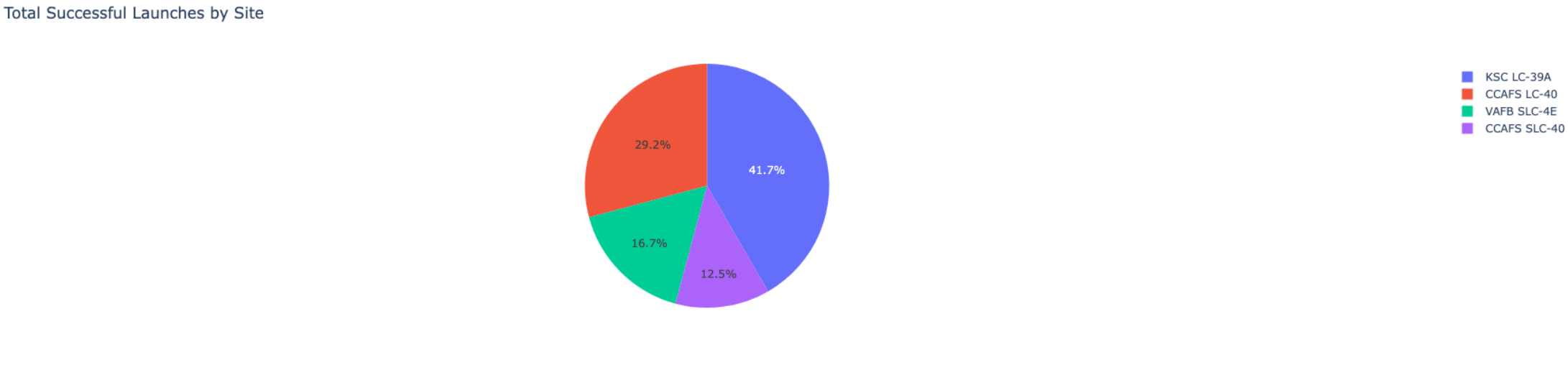


<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot

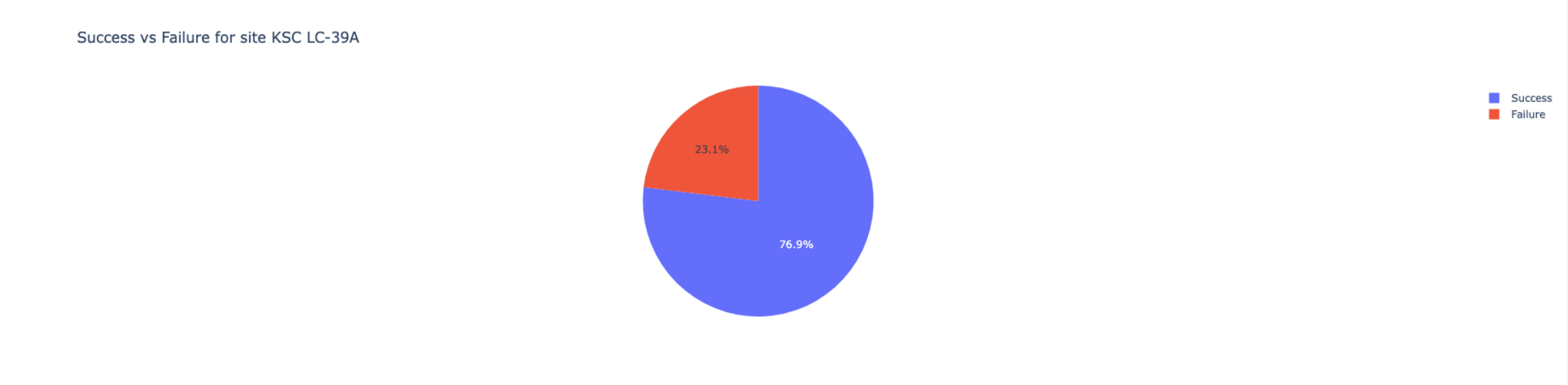

Section 4
Build a Dashboard with Plotly Dash

Successful Launches at Each Site



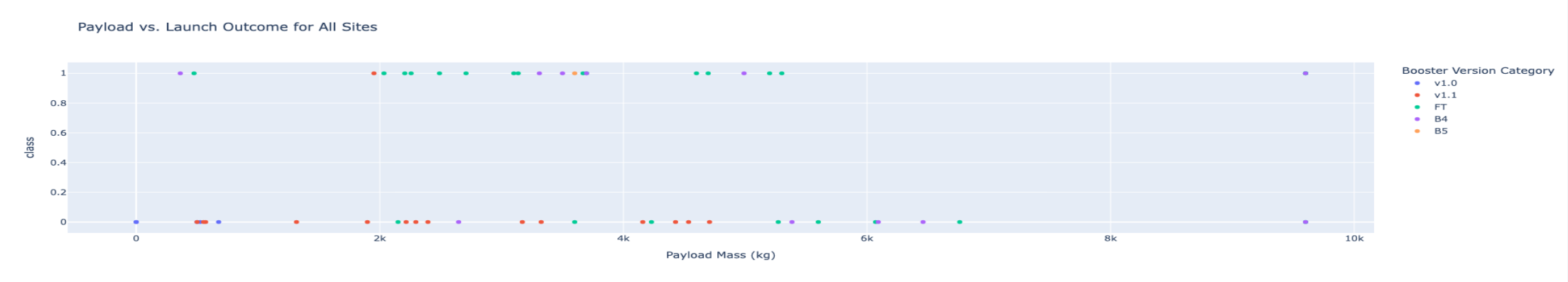
- The pie chart visually represents the **distribution of successful launches** across different SpaceX launch sites.
- Each slice corresponds to a specific launch site, and the **size of the slice** reflects the **proportion of successful launches** from that site.
- The chart includes **percent labels** for each site, providing a clear numerical comparison.
- **KSC LC-39A** accounts for the **largest share of successful launches** at 41.7%, indicating it is the most frequently used or most reliable site in this dataset.
- **CCAFS LC-40** also shows a significant share at 29.2%, making it another key launch location.
- **VAFB SLC-4E** and **CCAFS SLC-40** have smaller shares, reflecting either fewer launches or lower success counts.
- The chart highlights how launch activity is **concentrated around a few primary sites**, particularly KSC and CCAFS.
- This distribution gives insight into SpaceX's **operational preferences**, infrastructure usage, and historical success rates per site.

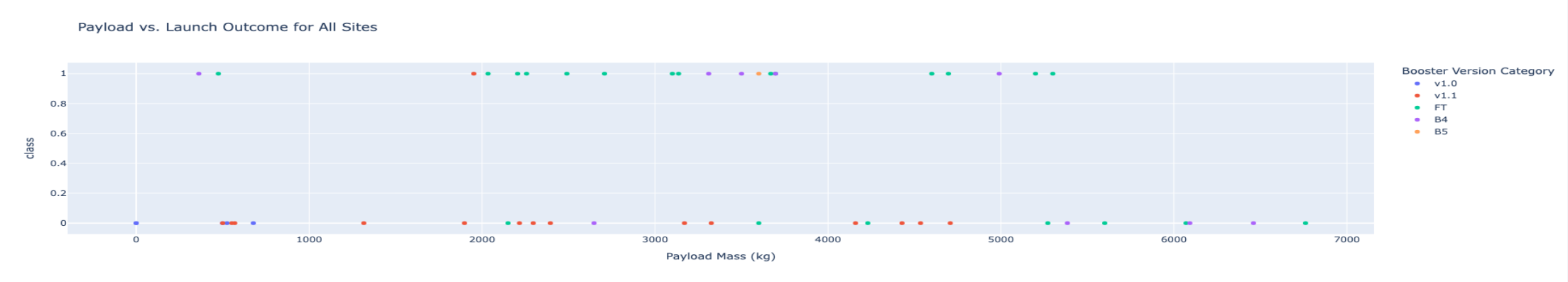
Most Successful Site



- The majority of launches from **KSC LC-39A have been successful**, indicating strong reliability for missions from this site.
- However, a **23.1% failure rate** is non-negligible and may reflect early-stage tests or challenges during specific launch campaigns.

Payload vs launch outcome Scatter Plots





- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

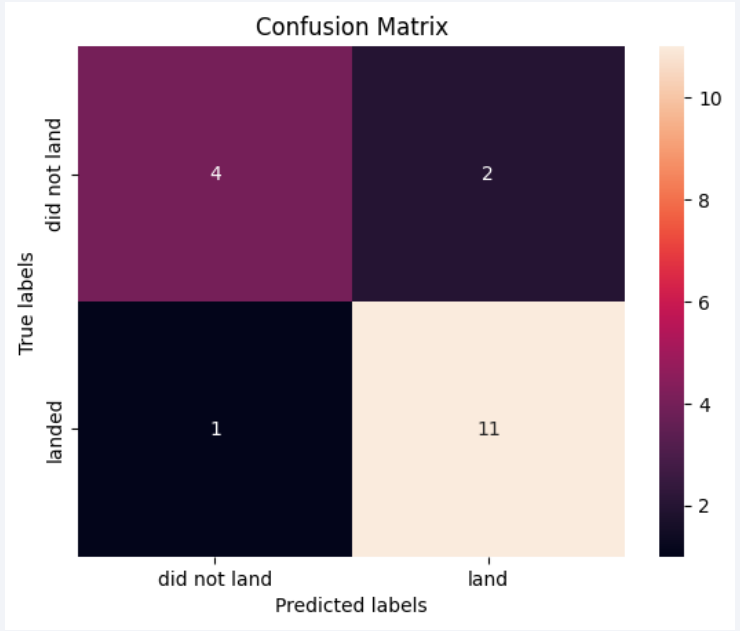
Classification Accuracy

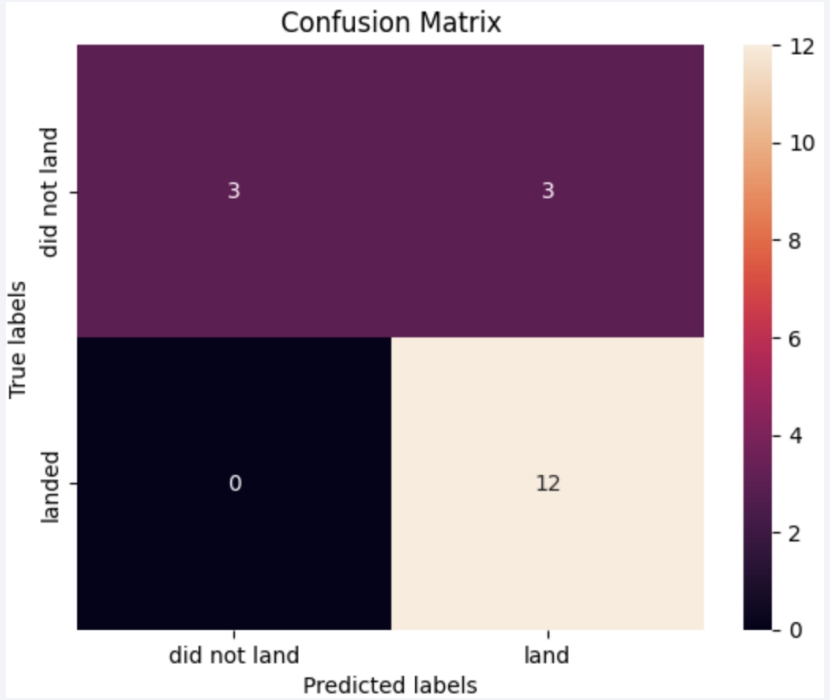
- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation
- I got that both models performed equally well. Decision Tree (left) and KNN (Right) both had an accuracy of 83%
-





Conclusions

- 1. **Exploratory data analysis revealed that launch site and payload mass significantly influence mission success**, with KSC LC-39A showing the highest success rate and certain payload ranges performing more reliably.
- 2. **Interactive visualizations and dashboard components (using Plotly Dash) enabled deeper insights**, allowing users to filter by launch site and payload range, which highlighted trends like higher success rates at specific sites and payloads.
- 3. **Machine learning models—including SVM and Logistic Regression—were able to predict launch success with up to 83% accuracy**, showing that SpaceX launch outcomes can be reasonably predicted using selected features such as site, orbit type, and booster version.
- 4. **The use of Folium for geospatial visualization successfully mapped global launch locations and outcomes**, helping to correlate geographic factors with mission results and giving a clear spatial perspective of SpaceX operations.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!