IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jesus Jimenez Martinez
January 19, 2025

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection: via API and Web Scraping
  - Data wrangling
  - Exploratory Data Analysis with SQL and Visualization
  - Interactive Visual Analytics with Folium. Interactive Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Data Insights:
    - Identified factors influencing Falcon 9 first stage landings.
    - Visualized geographical patterns and success rates.
  - Machine Learning Model Performance:
    - SVM, Logistic Regression and K-Nearest Neighbors have similar accuracy with Decision Tree being higher ~91% accuracy.
  - Key Findings:
    - Launch site and payload mass impact landing success.
    - Decision Tree model is the most effective predictor.

Introduction

Background and Context

- In this capstone project, our goal is to predict the successful landing of the Falcon 9 first stage. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars. Other providers cost goes up to165 million dollars each. Much of the savings is due to SpaceX's ability to reuse the first stage of the rocket. By predicting if the first stage will land, we can estimate launch costs and provide critical information to successfully bid against SpaceX for a rocket launch to companies like ours, SpaceY

Problem Statement

- Predict the successful landing of SpaceX's Falcon 9 first stage, a critical factor in determining the cost-efficiency of reusable rocket launches.

Why should we solve this problem?

- Solving this problem allows companies like SpaceY to make competitive bids against SpaceX by accurately estimating launch costs.

Questions to be answered?

- What is the historical success rate of Falcon 9 first stage landings?
- What factors/variables most significantly affects the success or failure of the first stage landing?
- Does the rate of successful landings increase over the years?
- Can we develop a predictive model that accurately forecasts the outcome of a landing?

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology**
  - Retrieval and consolidation from multiple SpaceX API
  - Web scaping tabular data from Wikipedia
- **Perform data wrangling**
  - Extracted relevant records
  - Flattened fields and resolved missing values
- **Perform exploratory data analysis (EDA) using visualization and SQL**
  - Visualize variable relationships
  - Look at the data in aggregate

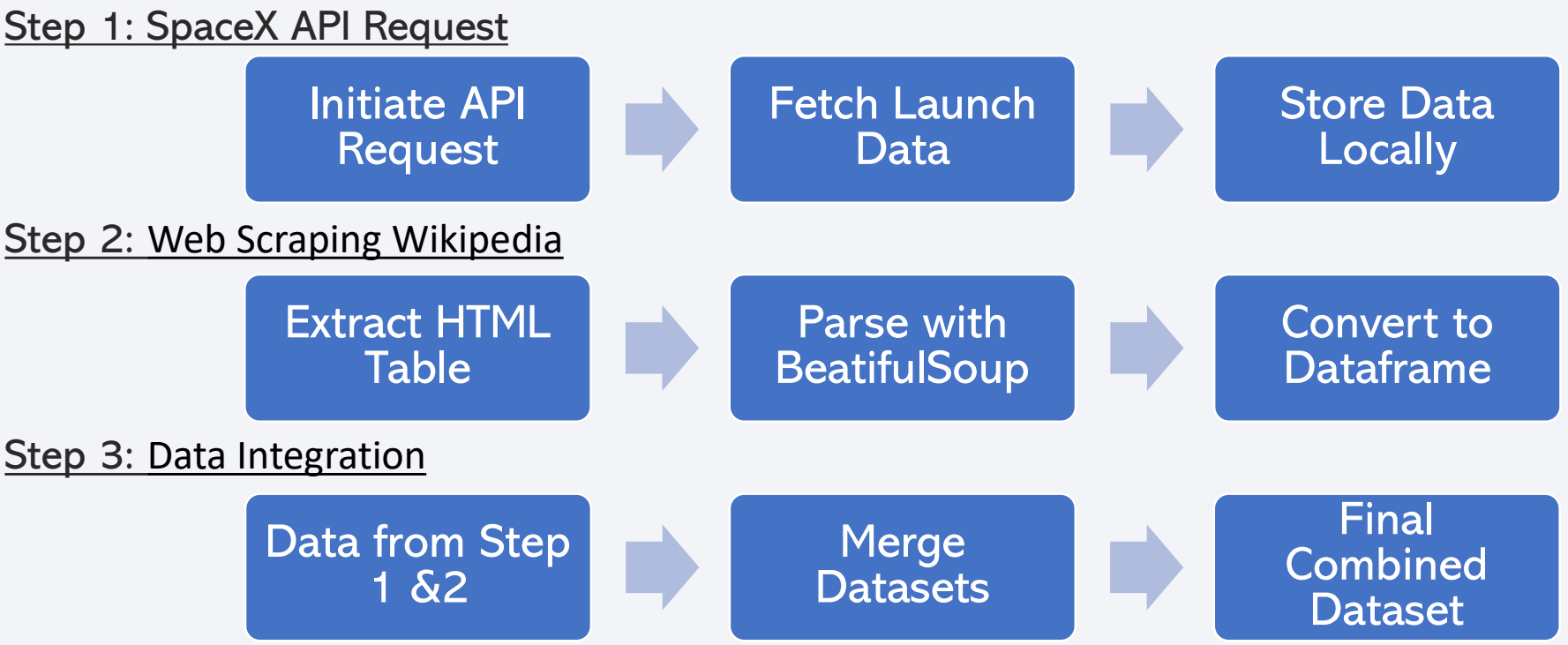
Methodology

Executive Summary (cont')

- **Perform interactive visual analytics using Folium and Plotly Dash**
  - Mark all launch sites on a map and successful and failed launches
  - Calculate distances to proximate locations
  - Provide for interactive exploration of the data
- **Perform predictive analysis using classification models**
  - Build, evaluate, and compare several predictive classification model

Data Collection

Data collection process combined API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry (from June 2021).Both steps of data collection methods were used to create a unique dataset with complete information about the launches for a more detailed analysis.



Data Collection – SpaceX API

Step 1: Initiate API Request

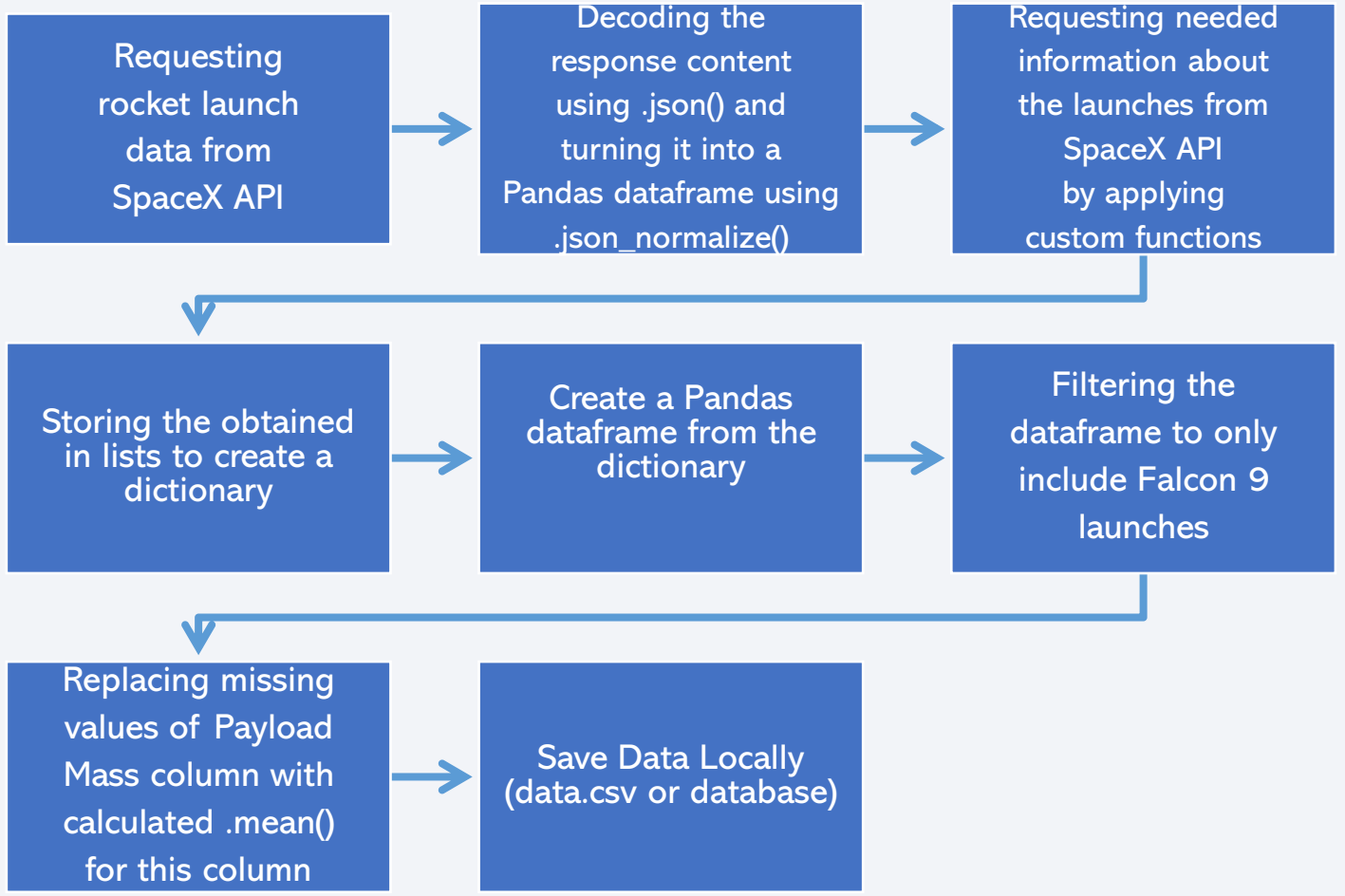
- Use Python's `requests` library to connect to the SpaceX API.
- Endpoint: `https://api.spacexdata.com/v4/launches`

Step 2: Parse API Response

- Convert API response from JSON to a Python dictionary.
- Extract relevant fields: launch date, launch site, payload mass, rocket type, outcome.

Step 3: Store Data Locally

- Save extracted data into a pandas Dataframe.
- Store the Dataframe locally for further processing.



GitHub URL: 1-Data Collection Api.ipynb

Data Collection - Scraping

Step 1: Initiate Web Scraping

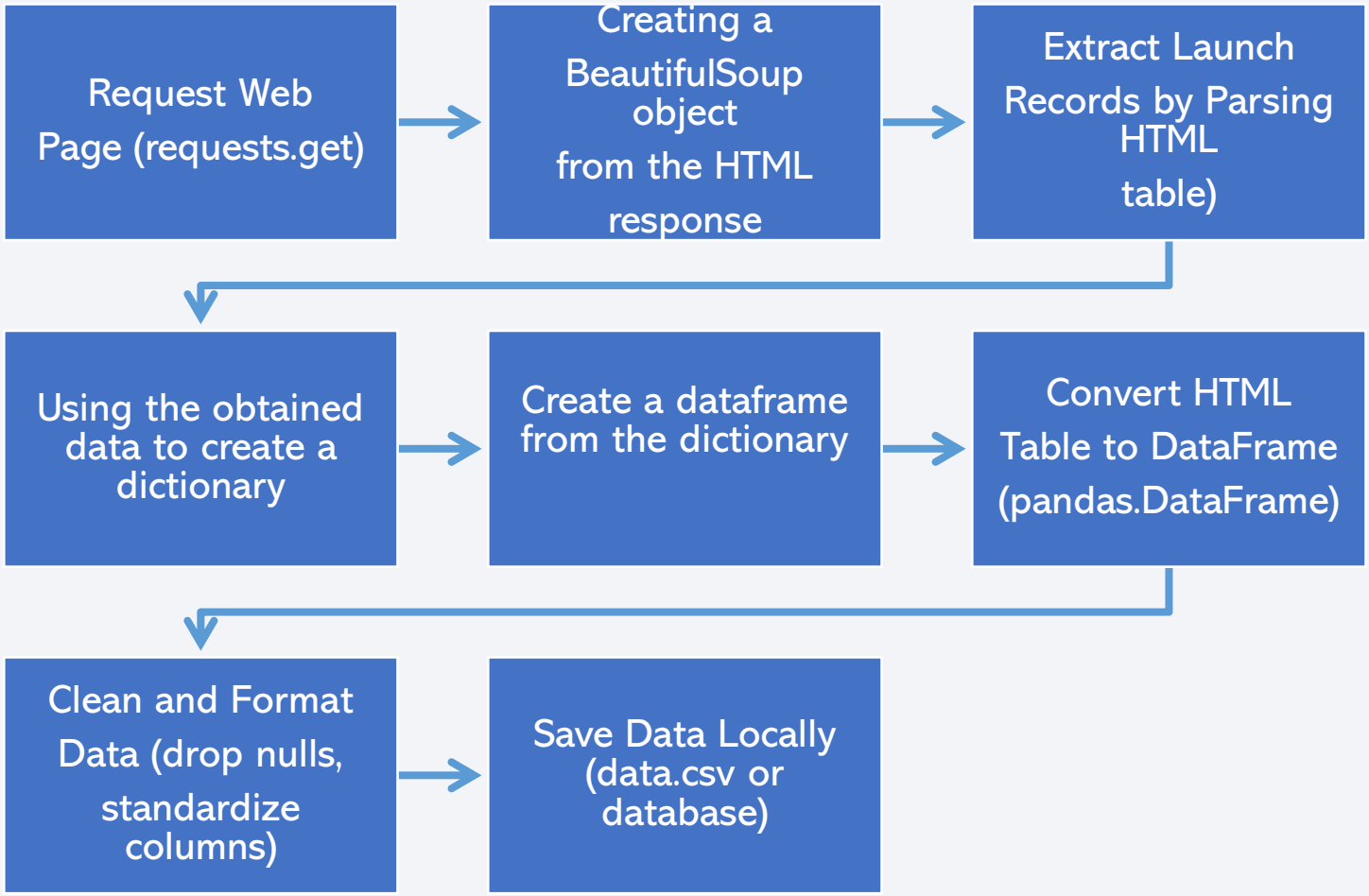
- Use Python's `requests` library to connect to the SpaceX API.
- Target URL: `https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches`

Step 2: Parse HTML Response

- Use `BeautifulSoup` to parse the HTML content.
- Extract the HTML table containing Falcon 9 launch records.

Step 3: Convert to Dataframe

- Convert the extracted HTML table into a pandas Dataframe.
- Store the Dataframe locally for further processing.



GitHub URL: 2-Data Collection Webscraping.ipynb

Data Wrangling

Initially some Exploratory Data Analysis (EDA) was performed on the dataset. Then Data wrangling involves several steps:

Step 1: Data Cleaning

- Identify and fill or remove missing values in the dataset.
- Use appropriate imputation techniques or drop rows/columns with excessive missing data

Step 2: Data Transformation

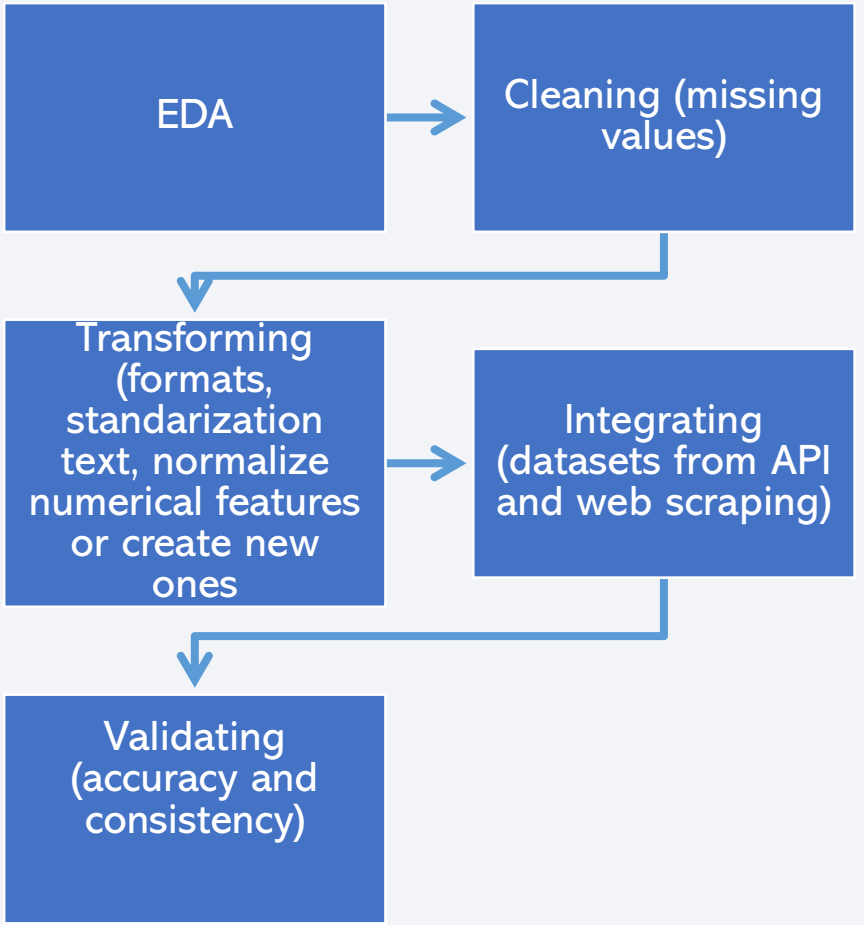
- Convert data types to appropriate formats.
- Standardize text.
- Create new features from existing data..
- Normalize/scale numerical features to ensure consistency.

Step 3: Data Integration

- Merge datasets collected from different sources (API, web scraping) into a single cohesive dataset.
- Ensure consistent column names and data formats across datasets.

Step 4: Data Validation

- Check for duplicate records and remove them.
- Verify the accuracy and consistency of data entries.

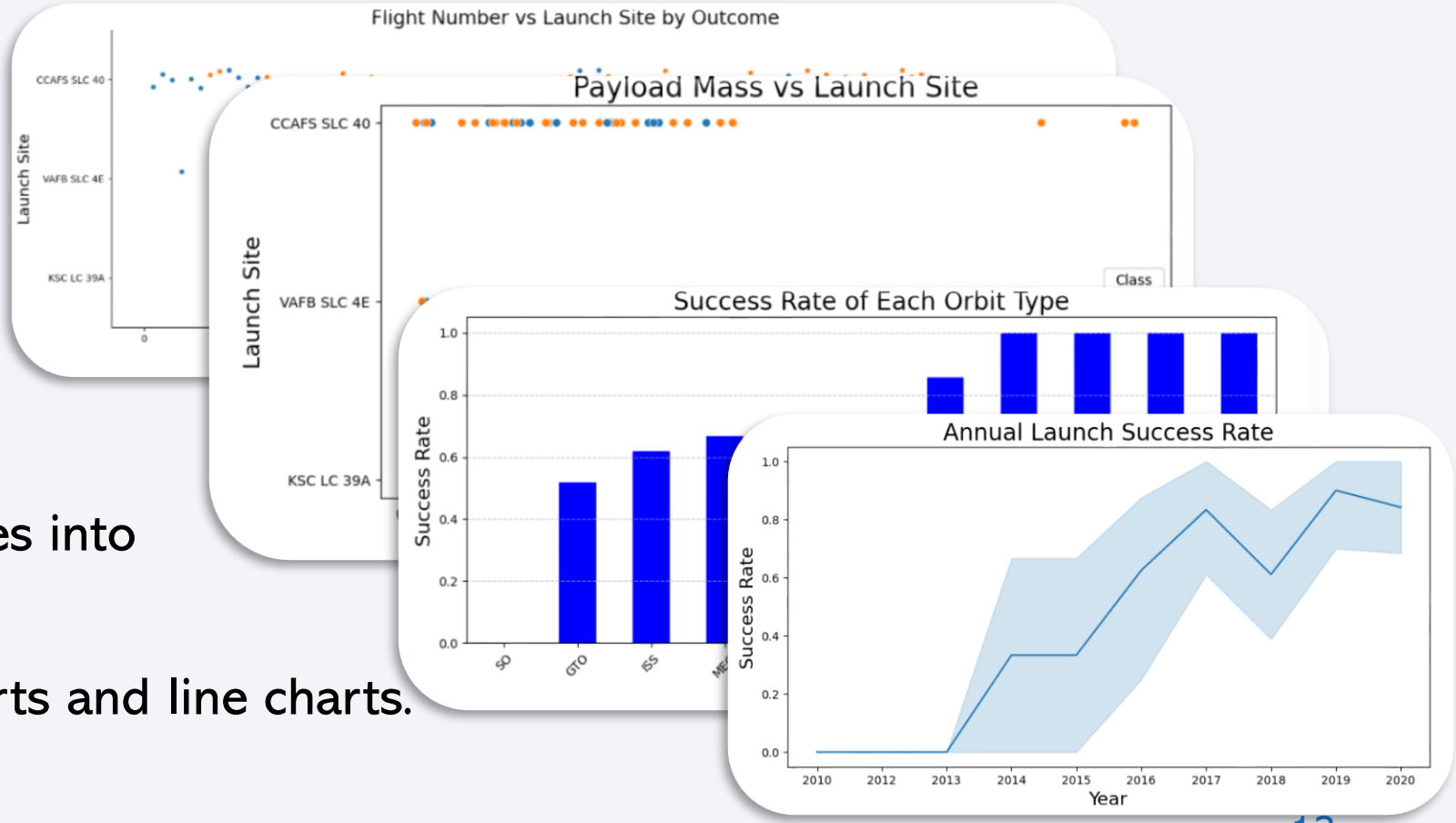


GitHub URL: 3-Data Wrangling.ipynb

EDA with Data Visualization

Visualize relationships to gain insight into the importance of each variable:

- Flight Number and Outcome
- Flight Number and Launch Site
- Payload and Launch Site
- Orbit and Outcome
- Flight Number and Orbit
- Payload and Orbit
- Yearly success rate



Transform categorical variables into "dummy" columns

Use of scatter plots , bar charts and line charts.

GitHub URL: 5-EDA_viz.ipynb

EDA with SQL

SQL Queries Performed:

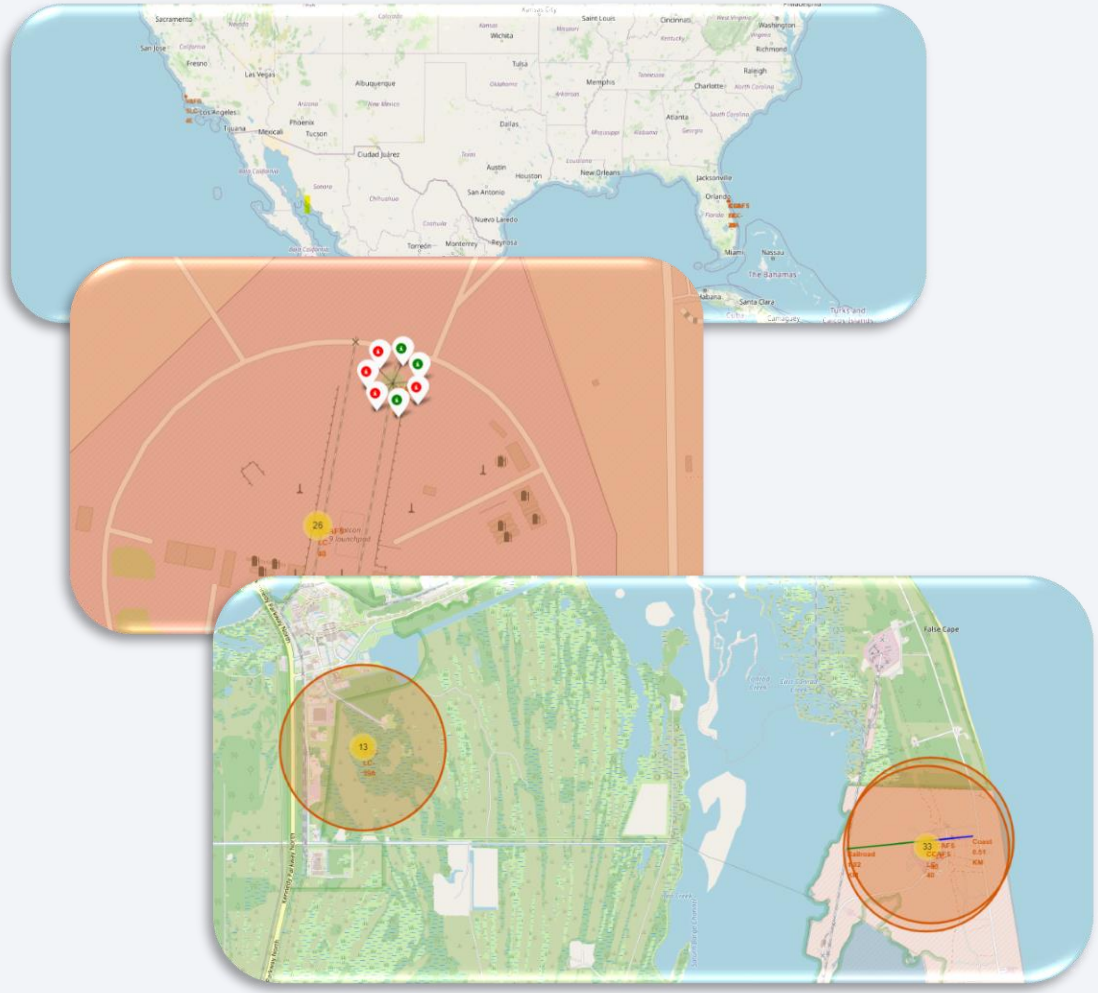
- Display the names of the unique launch sites in the space mission.
- Displaying 5 records where launch sites begin with the string 'CCA'.
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display the average payload mass carried by booster v1.1 Falcon 9.
- List the date of the first successful landing outcome in ground pad.
- List the booster versions with successful outcomes landing on the drone ship with payloads between 4000kg and 6000kg.
- List the total number of successful and failed mission outcomes
- List the names of all booster versions which carried the max payload mass
- List the records displaying month name, outcome, booster version, and launch site for missions with failure outcomes landing on a drone ship in 2015.
- Show the distribution of outcomes between June 4th, 2010 and March 20th, 2017 in descending order

GitHub URL: 4-EDA.ipynb

Build an Interactive Map with Folium

To find geographical patterns in the data the following items were marked on a map of launch sites:

- Markers of all Launch Sites (Circle, Popup Label and Text Label) using its latitude and longitude coordinates:
  - Johnson Space Center as a start location.
  - All Launch Sites to show their geographical locations
- Colored Markers of the launch outcomes for each Launch Site : success (Green) and failed (Red) launches; using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a launch site and proximate landmarks (e.g. coastline or railroads)

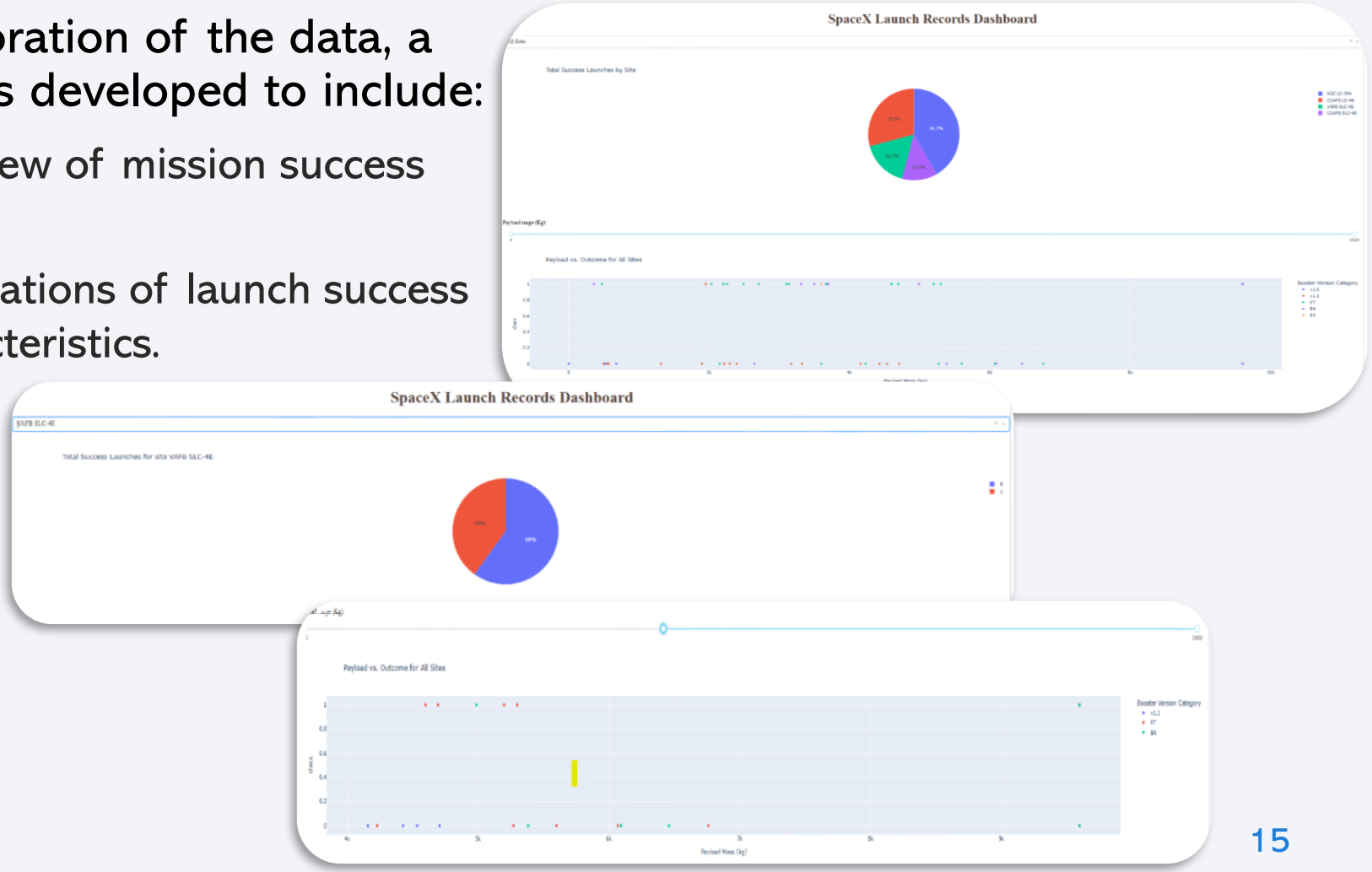


GitHub URL: 6- Interactive Viz-Launch Sites.ipynb

Build a Dashboard with Plotly Dash

To enable interactive exploration of the data, a Plotly Dash dashboard was developed to include:

- Pie charts providing overview of mission success rates per launch site.
- Scatter Plot showing correlations of launch success outcome vs payload characteristics.
- Filtering options:
  - Launch Site Dropdown
  - Range Slider for Payload

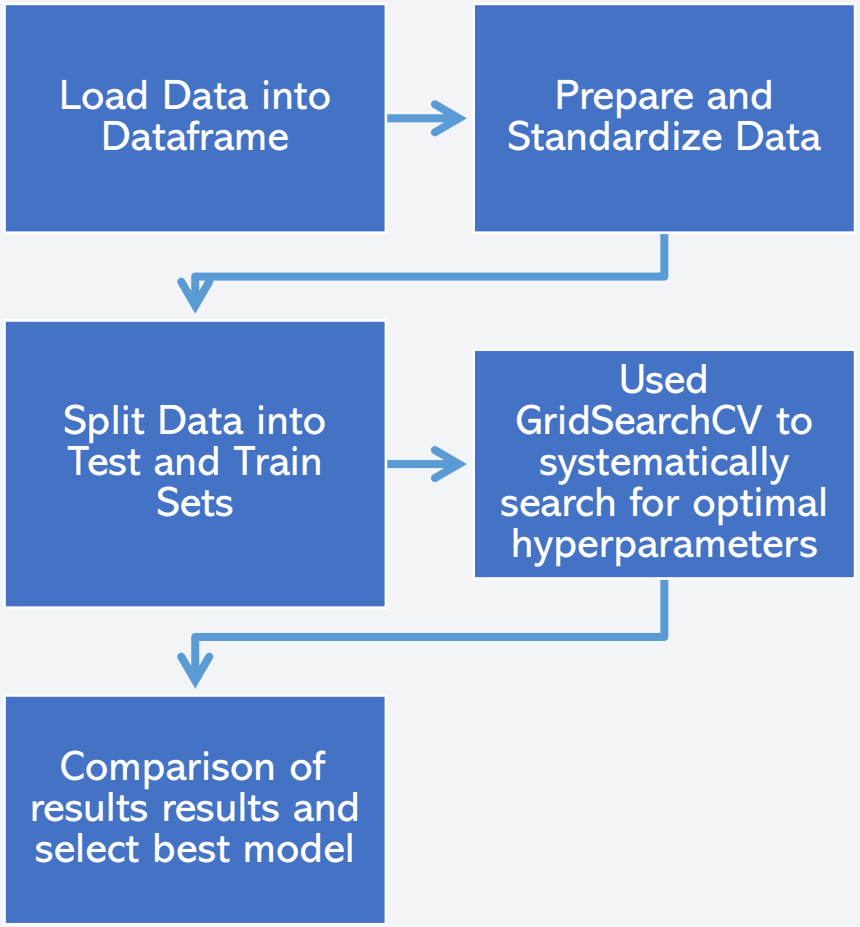


[GitHub URL: 7- Spacex Launch Sites Dashboard.py]

Predictive Analysis (Classification)

Four classification models were compared: logistic regression, support vector machine (SVM), decision tree and k nearest neighbors (KNN) and the following steps taken:

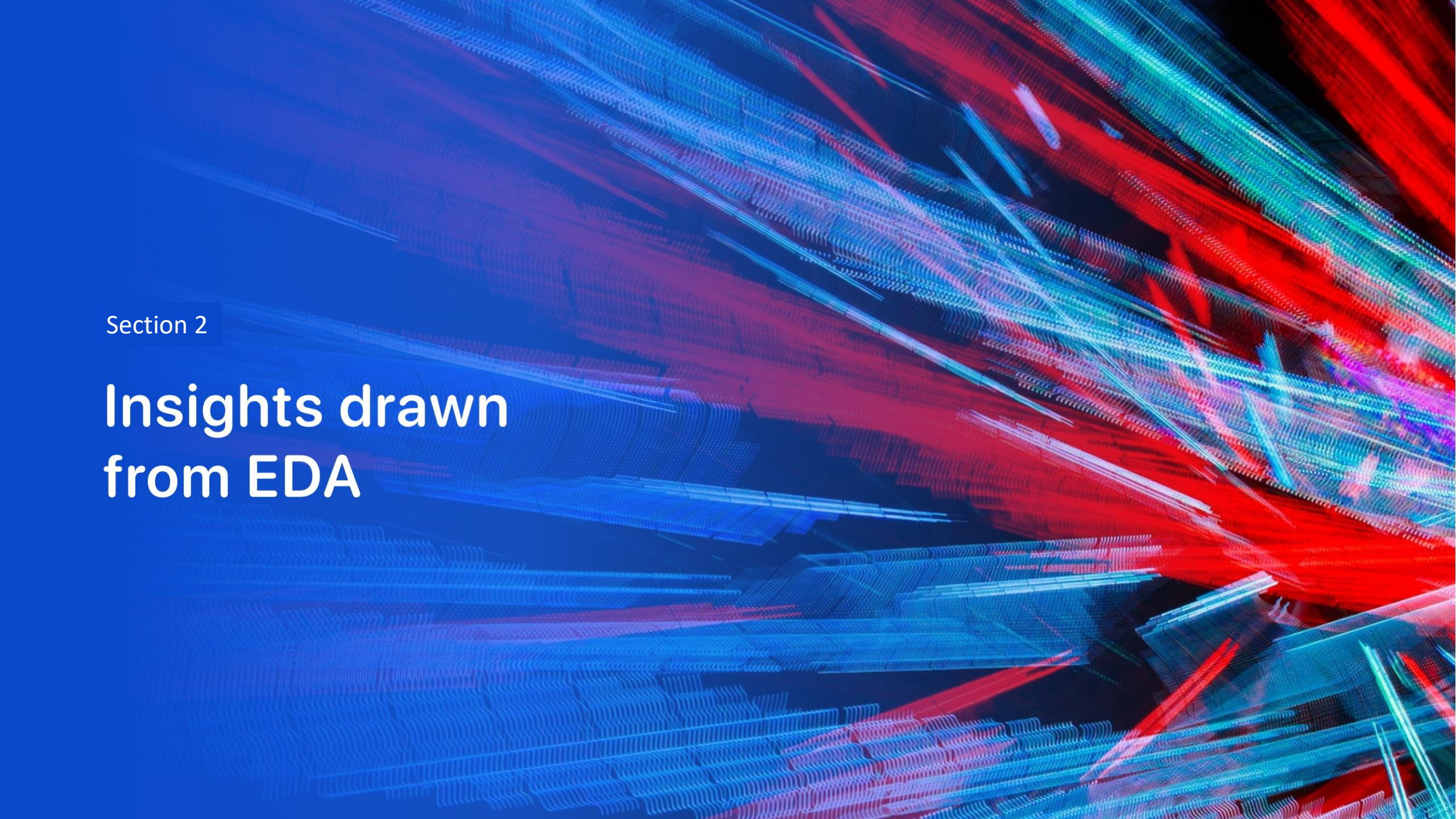
- Load data
- Apply StandardizedScaler on X
- Convert Y to numpy array
- Split training and testing data
- Use GridSearchCV to test hyperparameters for multiple algorithms
- Compared resulst and select best model



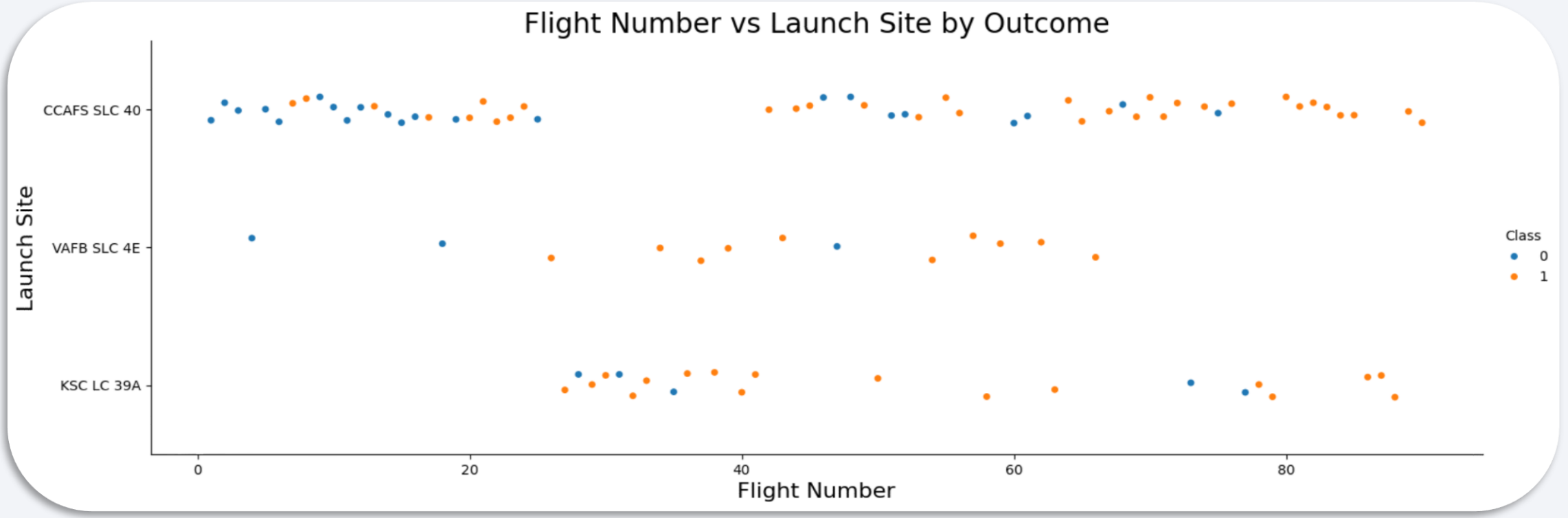
GitHub URL: 8-Machine Learning Prediction Models.ipynb

Results

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

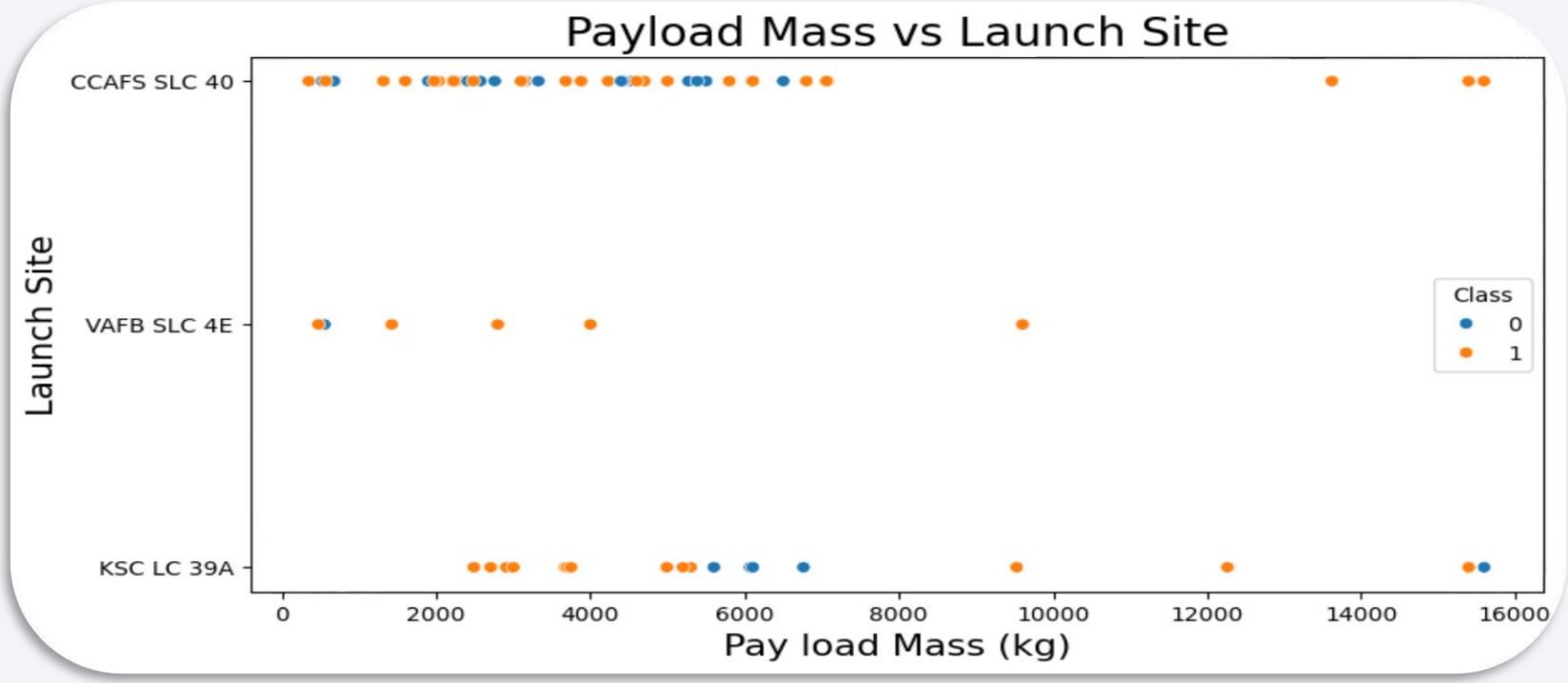

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Mix of first stage landing successes and failures for all major launch sites, with successes increasing over time (success outcome = 1 or orange). Early flights predominantly resulted in failures, indicating improvements to technology or process across the time.
- While CCAFS SLC 40 has the most total flights (~50%), VAFB SLC 4E and KSC LC 39A seem to have a relatively higher rate of successful landing outcomes

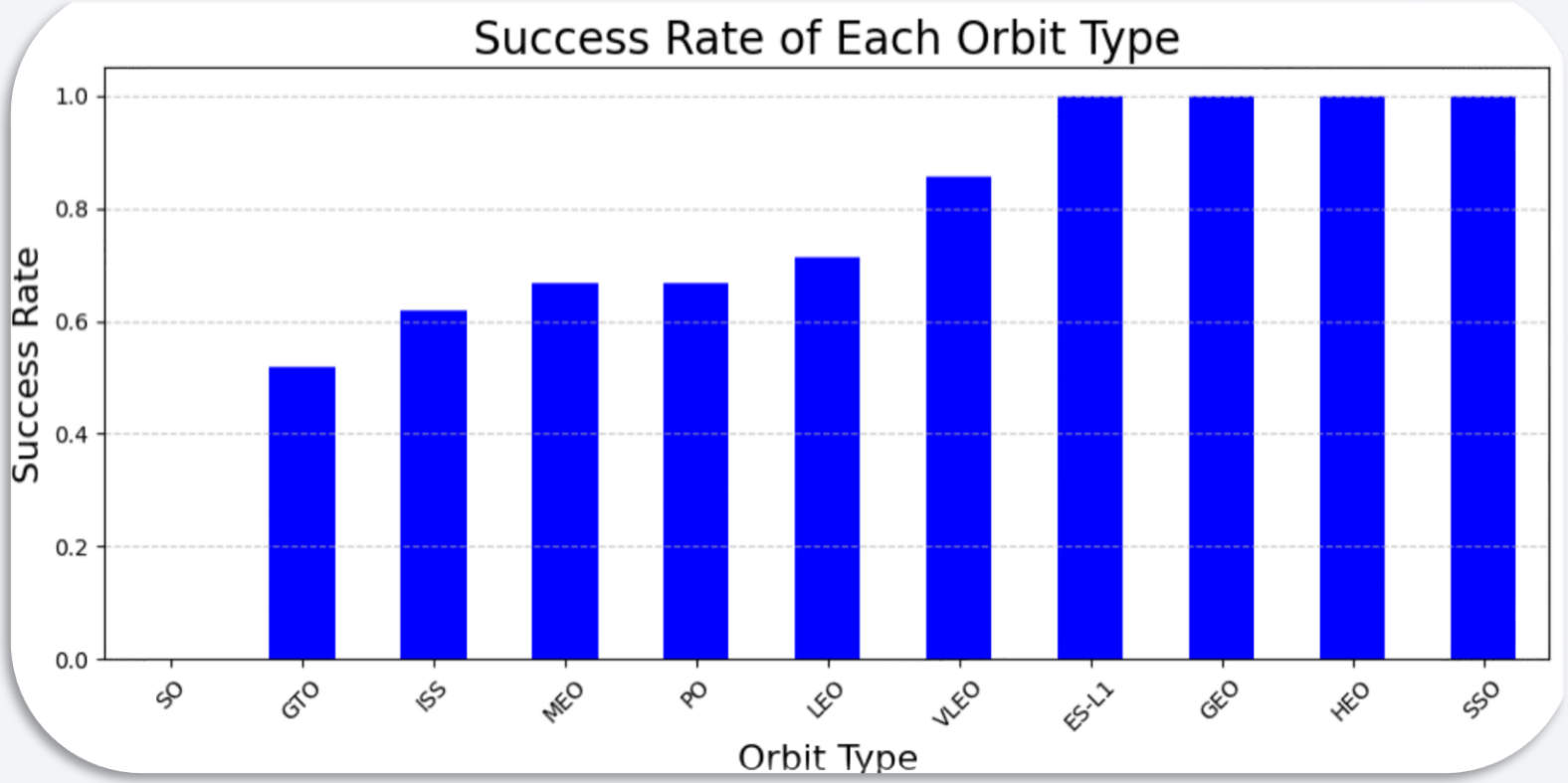
Payload vs. Launch Site



- All sites show a wide rage of payload weights. Most launches from the CCAFS SLC 40 and KSC LC 39A sites handle payloads below 8,000 kg, even if they have missions with heavy payloads (>12,000 kg) with a high rate of success. VAFB SLC 4E site doesn't launch heavy missions (>10,000 kg) but overall high success rate.
- Flights with lighter payloads (< 8,000 kg) represent the bulk of the landing failures suggesting that heavier payloads with a higher success rate, maybe have been influenced by technology improvements.

Success Rate vs. Orbit Type

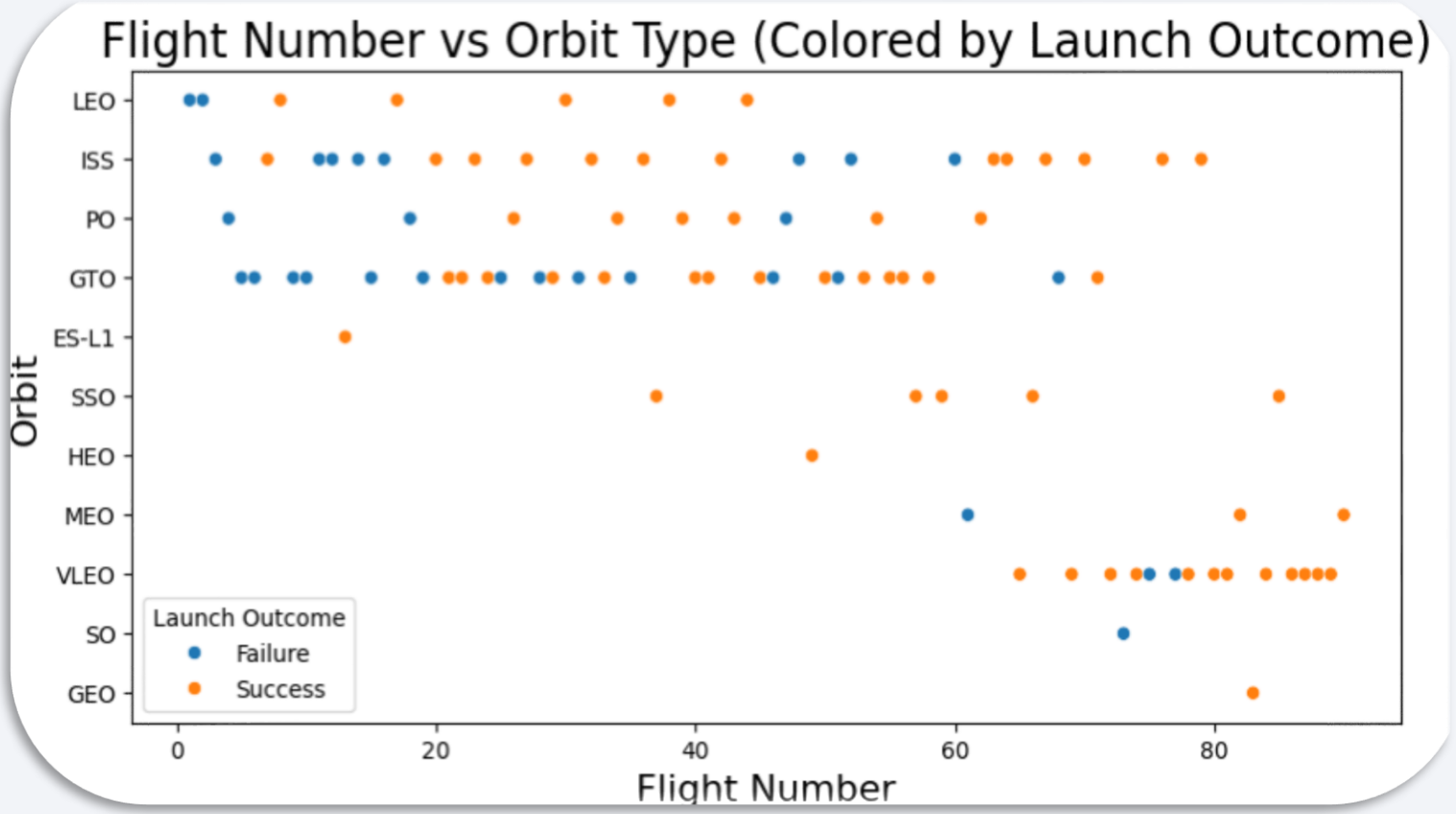
- Some orbits, such as ES-L1, SSO, HEO, and GEO consistently show high success rates. These orbits maybe be highly reliable for successful first stage landings.
- Others such as GTO, ISS, LEO, MEO and PO show more mixed outcomes, suggesting some orbits may introduce operational or technological challenges.



- Finally, there are orbits (i.e. SO) which don't have not enough data to provide an accurate analysis.

Flight Number vs. Orbit Type

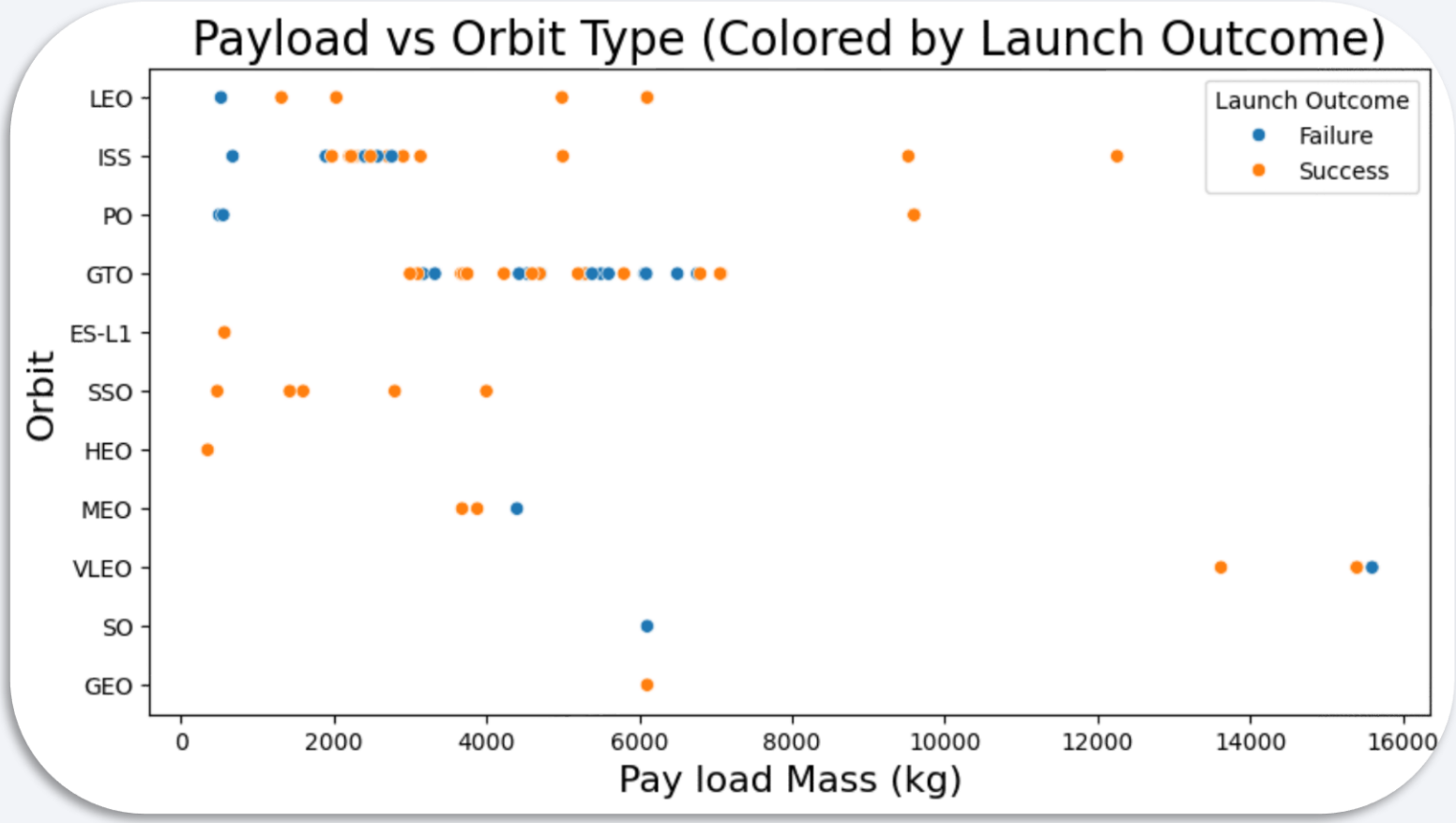
- Some orbits types (e.g. LEO, ISS, PO, GTO) have been launching missions from the beginning while others such as ES-L1, SSO, HEO, MEO, VLEO, SO and GEO have started later or even seemed to have stopped.
- Most of the orbits show an improvement in landing success rate over time due to most likely accumulation of operating experience or technology improvements.



- Orbits ISS, LEO, and VLEO perform especially better on landing success in their latest missions.

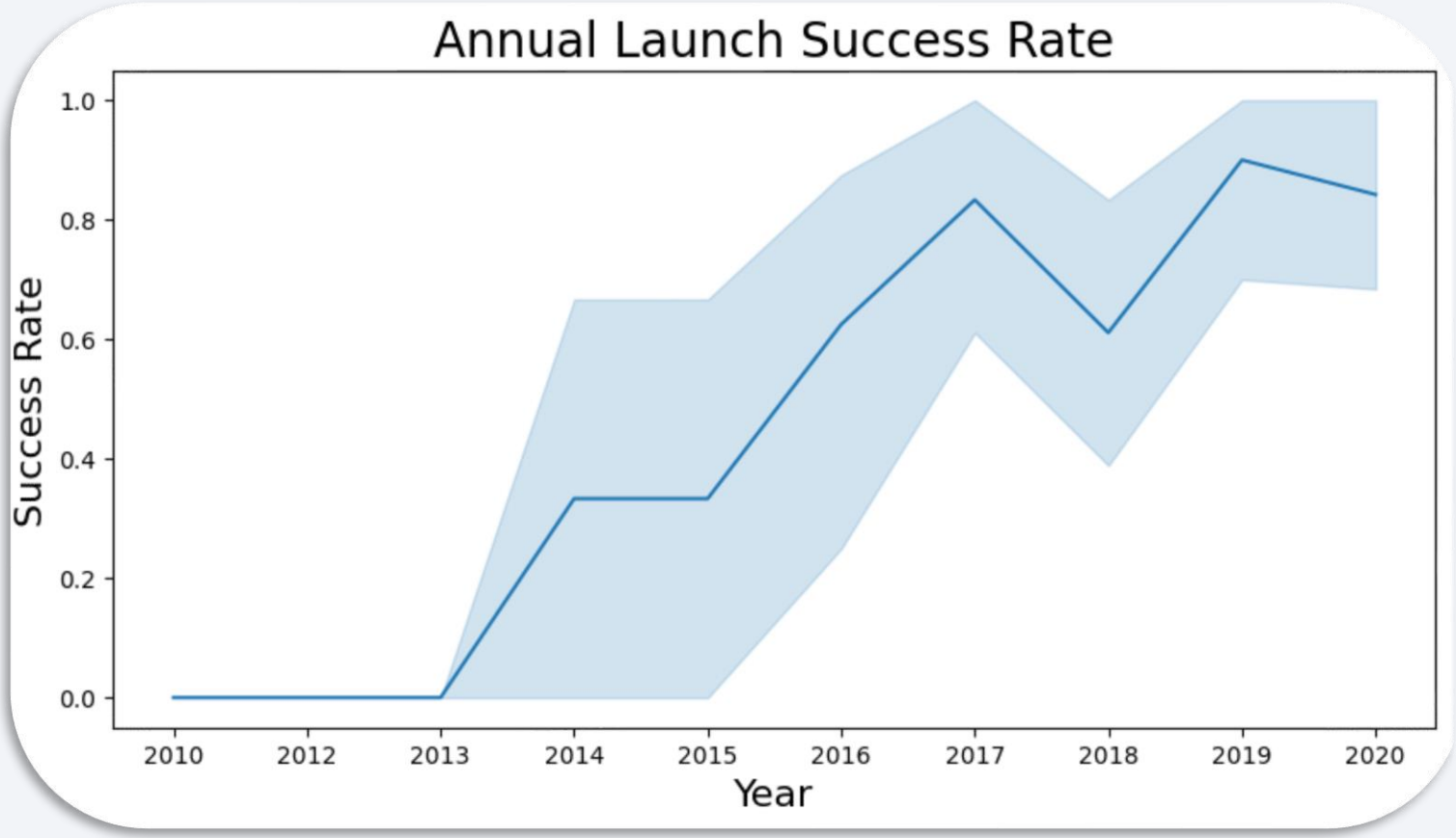
Payload vs. Orbit Type

- Some orbits , such as ISS, PO or LEO are represented across a wide range of payload masses, but others like SSO, MEO, HEO, VLEO and GEO show a generally lower range.
- Orbits with a constrained payload range, tend to show a higher rate of landing success as well as orbits using smaller payloads (<8,000 kg).
- Heavier payloads seem to have a negative influence on the orbits showing less and also mixed outcome results.



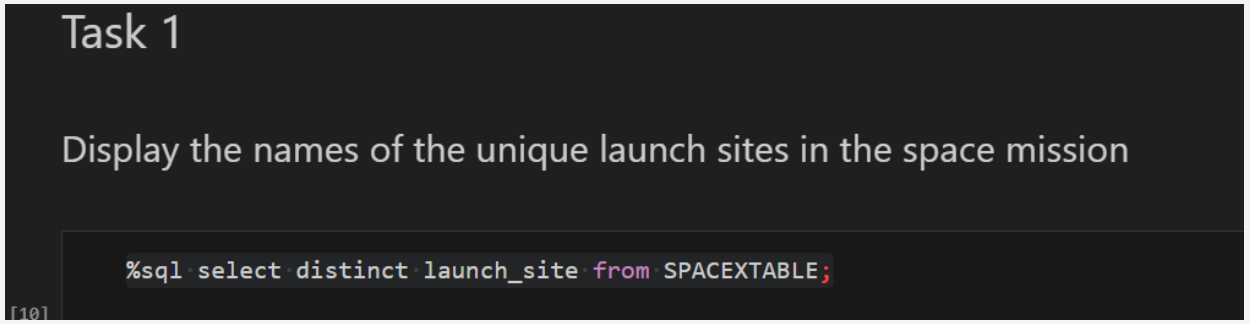
Launch Success Yearly Trend

- Yearly trend in success rate started continuously improving in 2013, reaching high reliability in first stage landings until 2020, with a minor setback in 2018.
- The first 3 years from 2010 until 2013 seem to be a learning period of adjustments and technology improvements.



All Launch Site Names

- There are four unique Launch Sites
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- Those values were obtained running the following SQL query, finding unique occurrences of “launch_site” values :



Launch Site Names Begin with 'CCA'

- The first 5 records where launch sites begin with `CCA` (Cape Canaveral):

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

- Those records were obtained running the following SQL query:

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

Total Payload Mass

- The total payload carried by boosters from NASA (CRS) is **45,596 kg.**
- That value was obtained running the following SQL query, summing all payloads whose codes whose codes ‘contain ‘CRS’, which corresponds to NASA:

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is **2,534.67 kg.**
- That value was obtained running the following SQL query, filtering data by the booster version F9 v1.1 and calculating the average payload mass:

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was: **2015-12-22**
- That value was obtained running the following SQL query, filtering data by successful landing outcome on ground pad and getting the minimum value for date:

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

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

- Those values were obtained running the following SQL query, selecting distinct booster versions according to the filters above.

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

Mission_Outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Those values were obtained running the following SQL query, grouping mission outcomes and counting records for each group.

Task 7

List the total number of successful and failure mission outcomes

+ Code + Markdown

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

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

- Those values were obtained running the following SQL query.

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

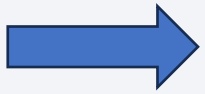
```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

2015 Launch Records

- The lis of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 is:

Month	Landing_Outcome	Booster_Version	Launch_Site	Date
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- Those values were obtained running the following SQL query.



```
%%sql

SELECT
    CASE strftime('%m', Date)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END as Month,
    Landing_Outcome, Booster_Version, Launch_Site, Date
FROM SPACEXTABLE
WHERE strftime('%Y', Date) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

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

- The ranking 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:

Landing_Outcome	count_outcomes
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

- Those values were obtained running the following SQL query.

Task 10

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.

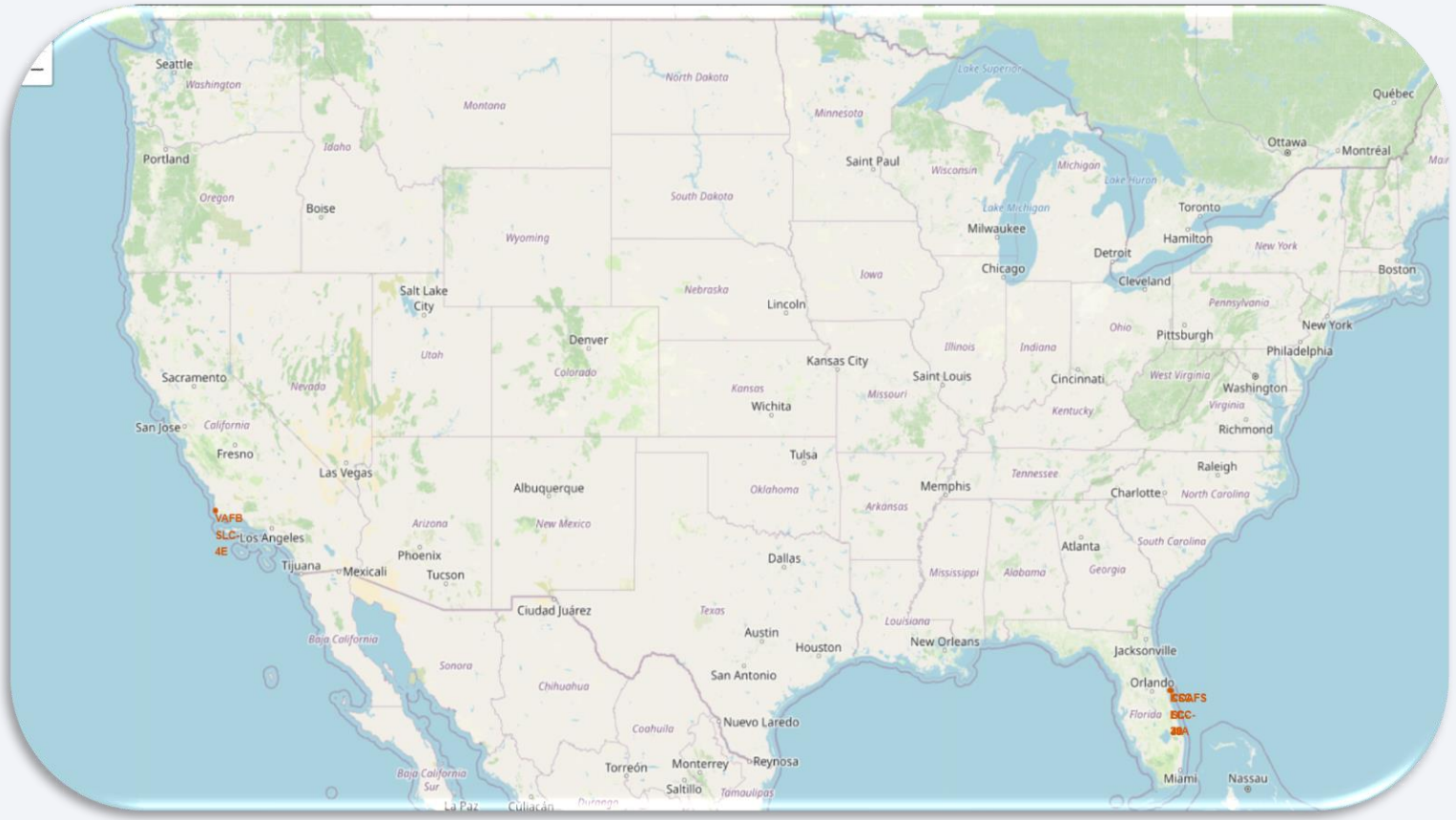
```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

Section 3

Launch Sites Proximities Analysis

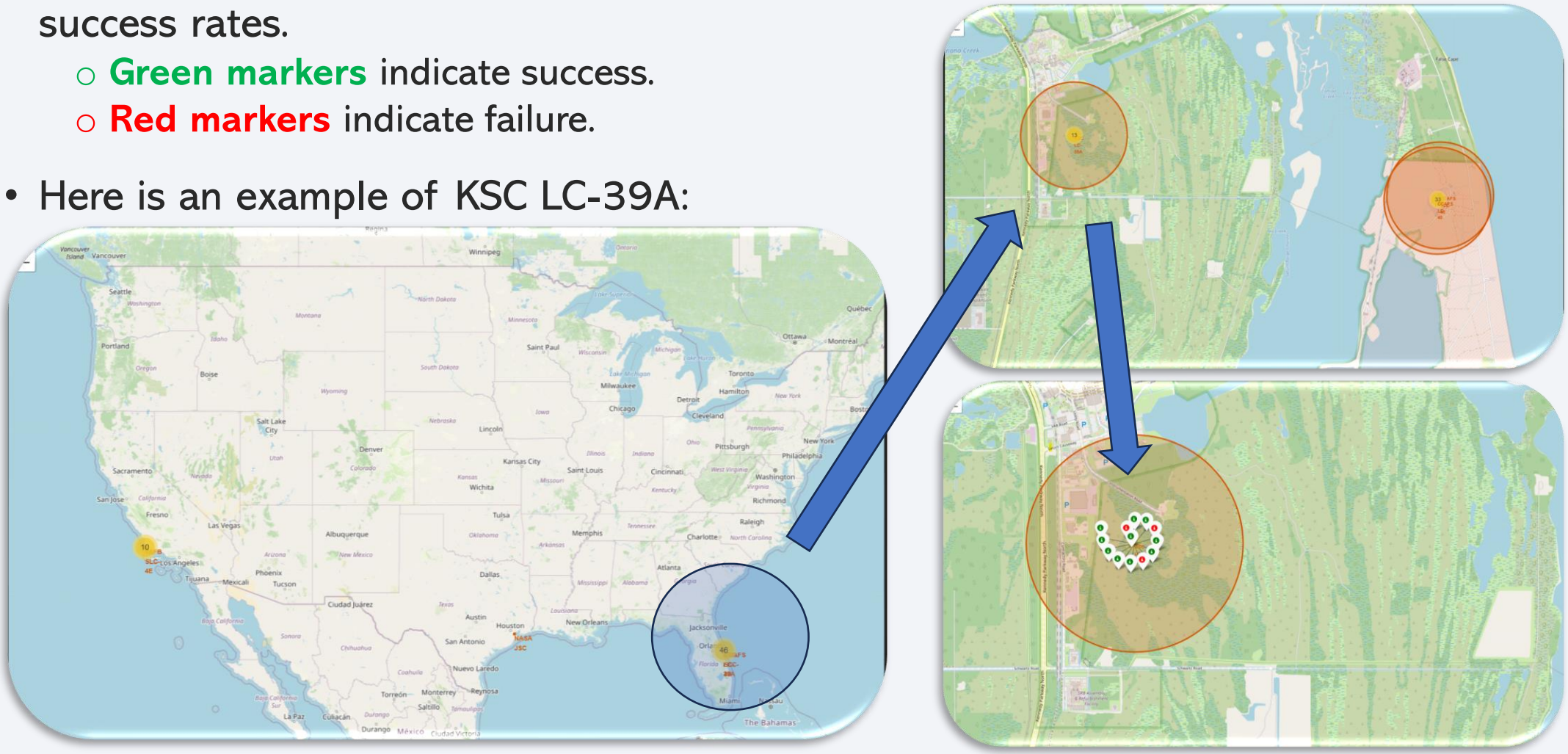
All Launch Sites

- All launch sites are located near coastal regions (i.e. Florida and California) to reduce risk of catastrophic failures affecting human safety (due to debris dropping or exploding) and close to the Equator.
- Also, launch sites are near to transportation infrastructures like roads or railroads for logistic advantages.



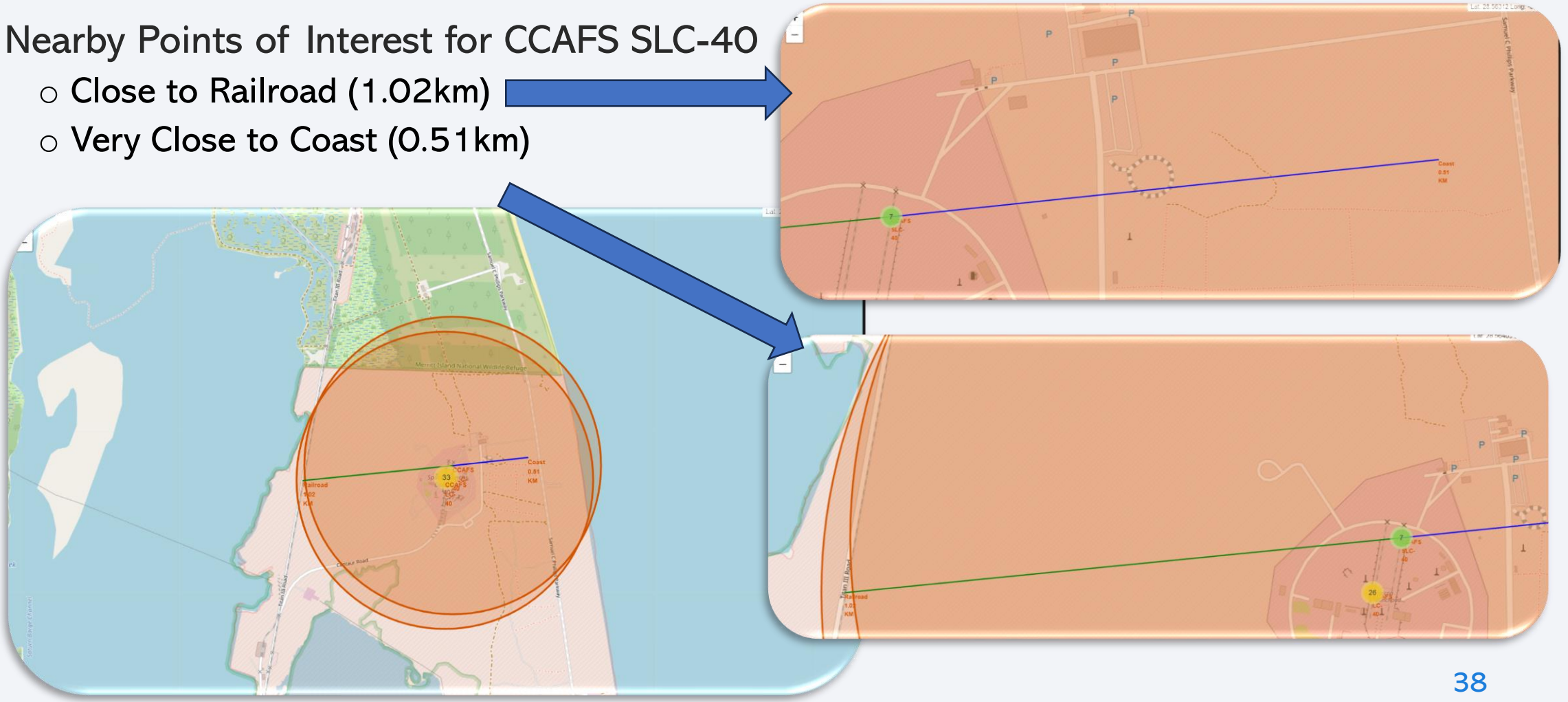
Launch Outcomes by Site

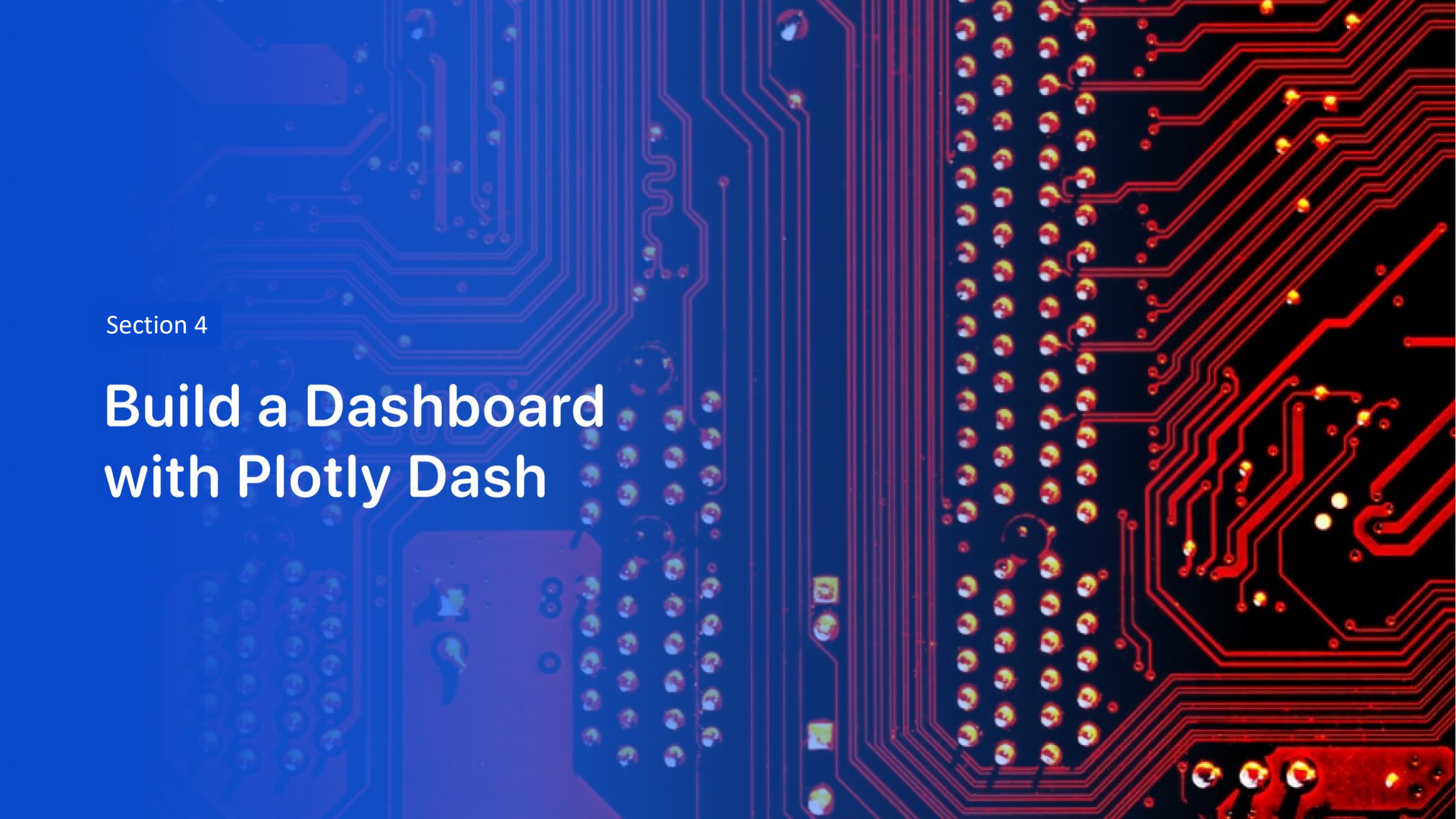
- Using colour-labeled markers we can easily identify which launch sites have higher success rates.
  - **Green markers** indicate success.
  - **Red markers** indicate failure.
- Here is an example of KSC LC-39A:



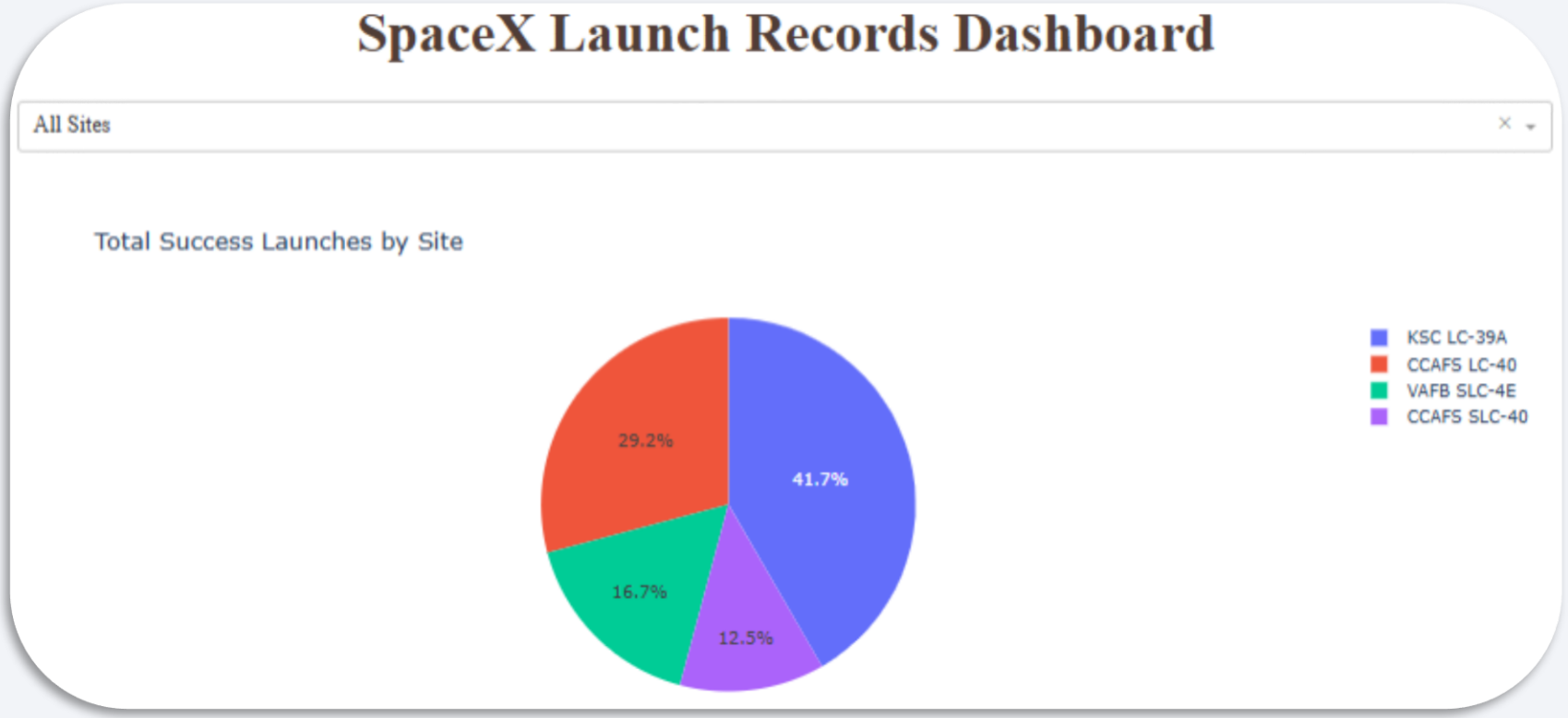
Proximate Points of Interest

- Nearby Points of Interest for CCAFS SLC-40
  - Close to Railroad (1.02km)
  - Very Close to Coast (0.51km)



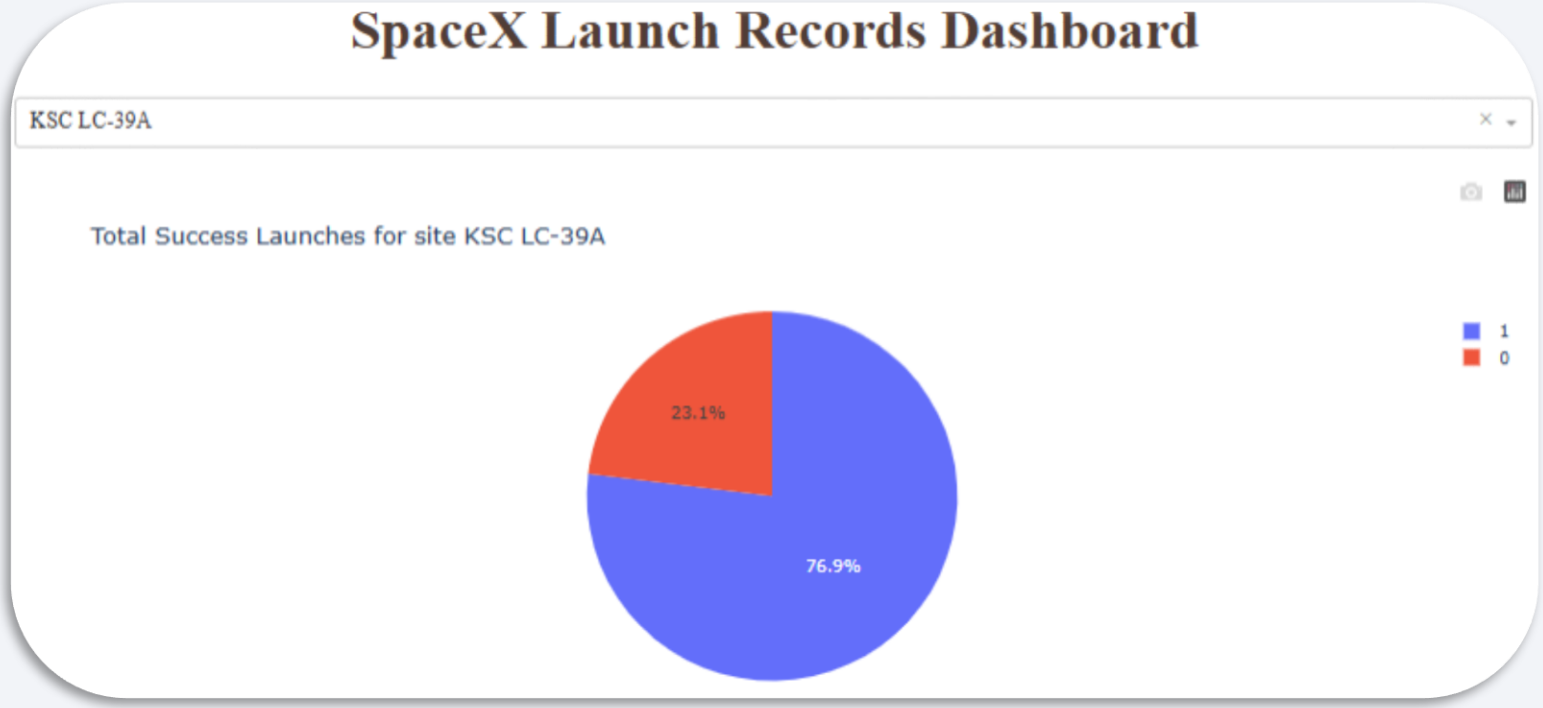

Section 4
Build a Dashboard with Plotly Dash

Launching Success by Site



- KSC LC-39A experienced the highest ratio of successful landings (41.7%), followed by CCAFS LC-40 (29.2%). Being VAFB SLC-4E (16.7%) and CCAFS SLC-40 (12.5%) the lowest.
- KSC LC-39A seems to be the launch site more reliable for SpaceX launches.

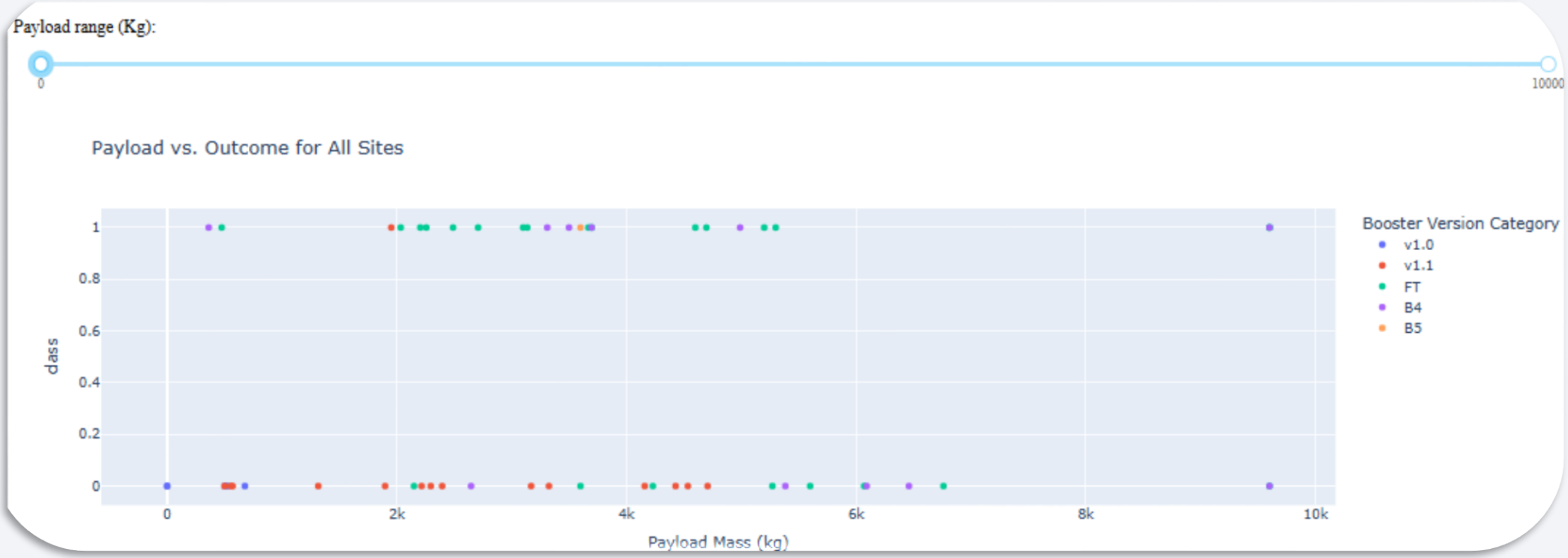
Per-site Launch Success Ratio: KSC LC-39A



- KSC LC-39A experienced the highest ratio of successful landings with 41.7%) proving its reliability and effectiveness as a launch site.

NOTE: Class =1 means successful landing.

Payload Mass Range vs Launch Success



- Payload masses under 6,000kg and FT boosters are the most successful combination. On the other hand, v1.1 boosters performed the worst within the same range. FT abd V1.1 seem to be the most used booster types.
- Over 6,000kg payload mass, neither B4 or FT booster are successful.

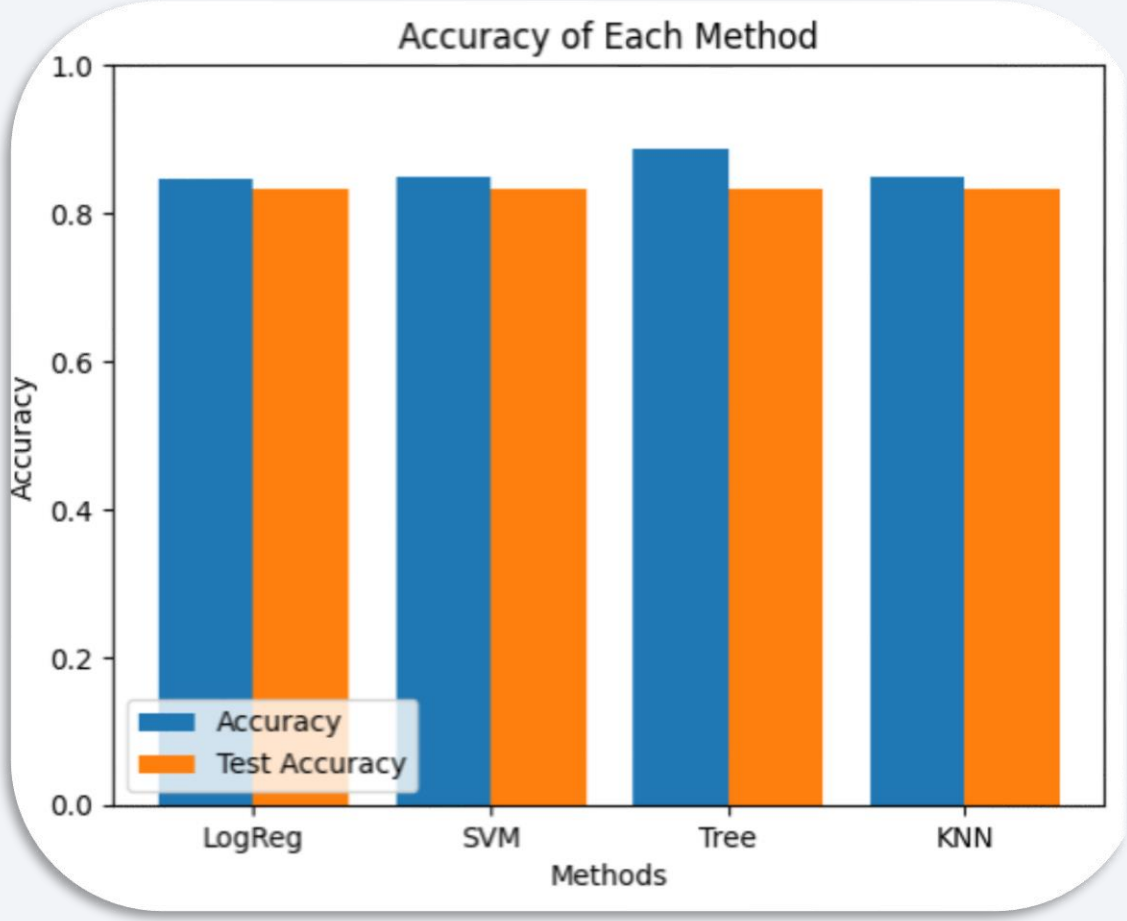
Section 5

Predictive Analysis (Classification)

Classification Accuracy

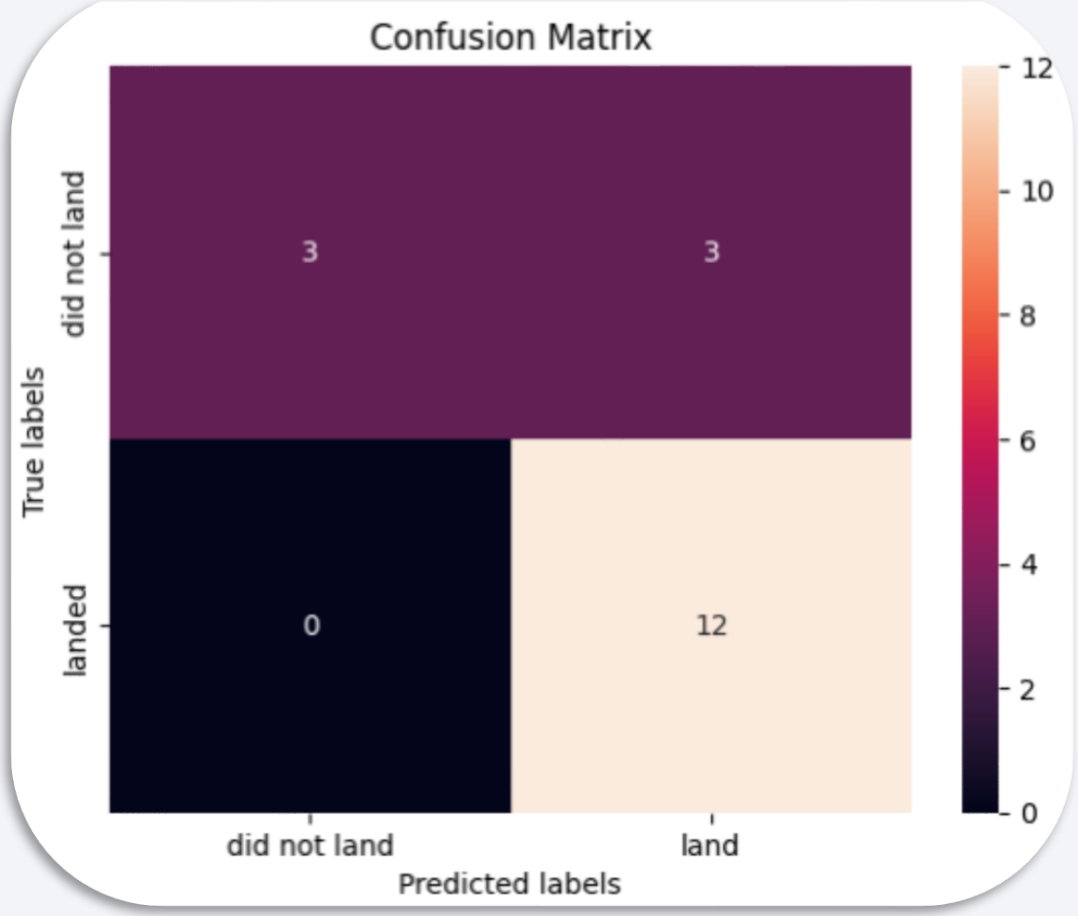
- Four machine learning models were created : logistic regression, support vector machine (SVM), decision tree and k nearest neighbors (KNN). Accuracy per the test data is the same in all models but if we examine the model using the whole dataset, we can see a difference.
- DecisionTreeClassifier was the most accurate

	LogReg	SVM	Tree	KNN
Accuracy	0.866667	0.877778	0.911111	0.855556



Confusion Matrix

- In the confusion matrix of the Decision Tree Classifier model we can observe:
  - 12 observations were correctly predicted as successful landings (true positive)
  - 3 observations were correctly predicted as failed landings (true negative)
  - 3 observation was incorrectly predicted as a successful landing (false positive)
  - No observations were incorrectly predicted as a failed landing (false negative)

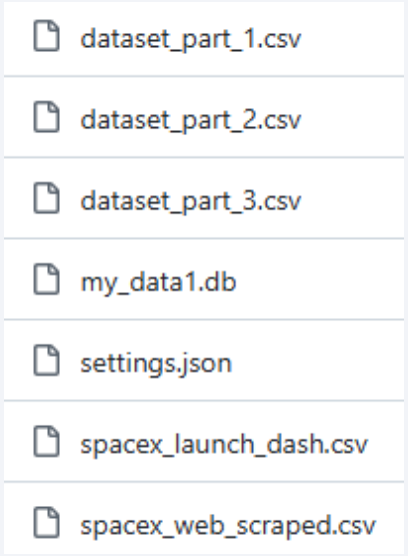


Conclusions

- Success rates increase over time, across all analyzed factors, which seems to be caused by continuous and incremental operation experience and technological developments.
- Different orbits have varying success rates, with ES-L1, SSO, HEO, and GEO showing consistently successful outcomes.
- Launches with a low payload mass show better results than launches with a larger payload mass. FT booster version has a high success rate across various payload masses, proving its reliability and robustness compared to other booster versions.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- Launch site KSC LC-39A with a highly predictive factor seems to be the top performer.
- The 4 predictive models evaluated were able to predict landing outcome with an acceptable level of accuracy. Decision Tree Classifier produced the best results with high accuracy so it can be used to predict successful landings and increase profits.

Appendix

- Data Sources
  - SpaceX API
  - Wikipedia List of Falcon 9 and Falcon Heavy launches (June 2021)
- Other files used with the Jupyper Notebooks to complete this project are uploaded in GitHub.



GitHub URL: IBM-Applied-Data-Science-Capstone-Project

Thank you!