IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abhishek Santosh Gaikwad
08-04-2025

Outline

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

Executive Summary

Summary of methodologies:

- **Data Collection**: Web scraping and SpaceX API
- **Data Wrangling**: Cleaning and preparing data for analysis
- **Exploratory Data Analysis (EDA)**:
  - Data visualization for insights
  - SQL queries for deeper analysis
- **Interactive Tools**:
  - Folium for building maps
  - Plotly Dash for interactive dashboards
- **Predictive Analysis**: Machine learning classification models to predict launch success

Summary of all results:

- Successfully collected and analyzed data from public sources.
- EDA identified key features influencing launch success.
- Machine learning models provided insights into important launch characteristics.
- Interactive analytics helped visualize and interpret findings effectively.

Introduction

Space Y wants to compete with SpaceX in the space industry. To succeed, it needs to understand what makes rocket landings successful and how to reduce launch costs.

Problems to Solve

- ➔ How can we predict successful rocket landings to estimate launch costs?
- ➔ What is the best location for rocket launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources
  - ➔ SpaceX API: Used to fetch past launch data from (https://api.spacexdata.com/v4/launches/past.)
  - ➔ Wikipedia Snapshot: Scraped historical launch data from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - The data was analyzed using Pandas to examine launch sites, orbit frequencies, and mission outcomes while also classifying landing results. Furthermore, new labels were generated to provide clearer insights into landing outcomes.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

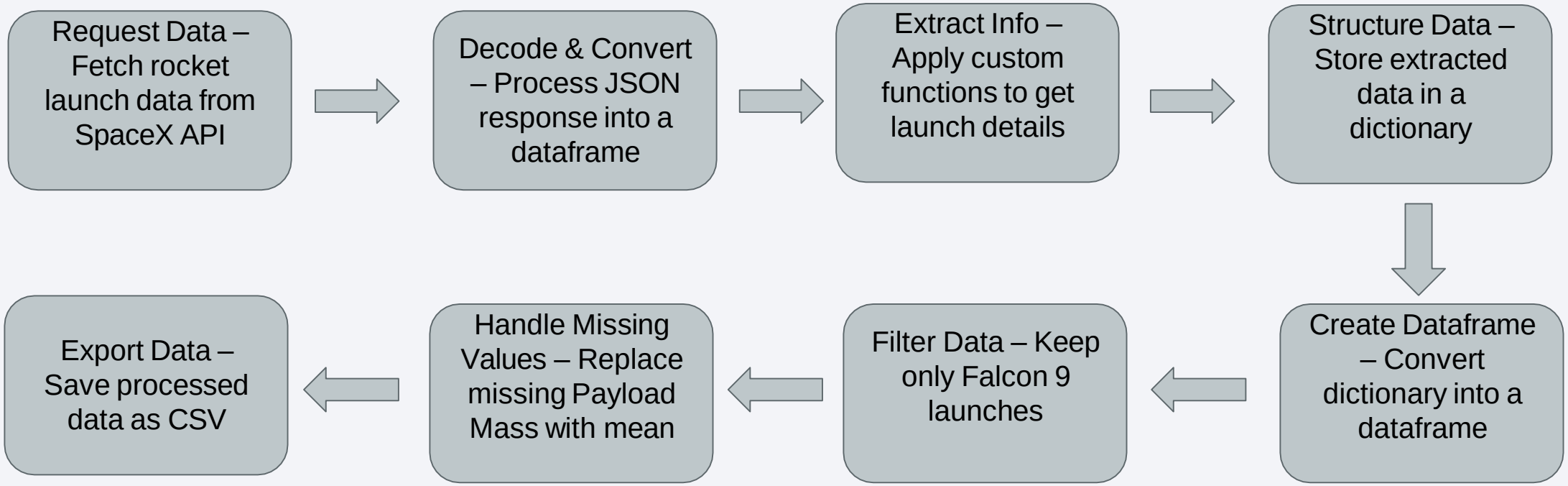
- Perform predictive analysis using classification models

  The collected data was **normalized**, split into **training and test sets**, and evaluated using four classification models. Each model's accuracy was tested with different parameter combinations.

Data Collection

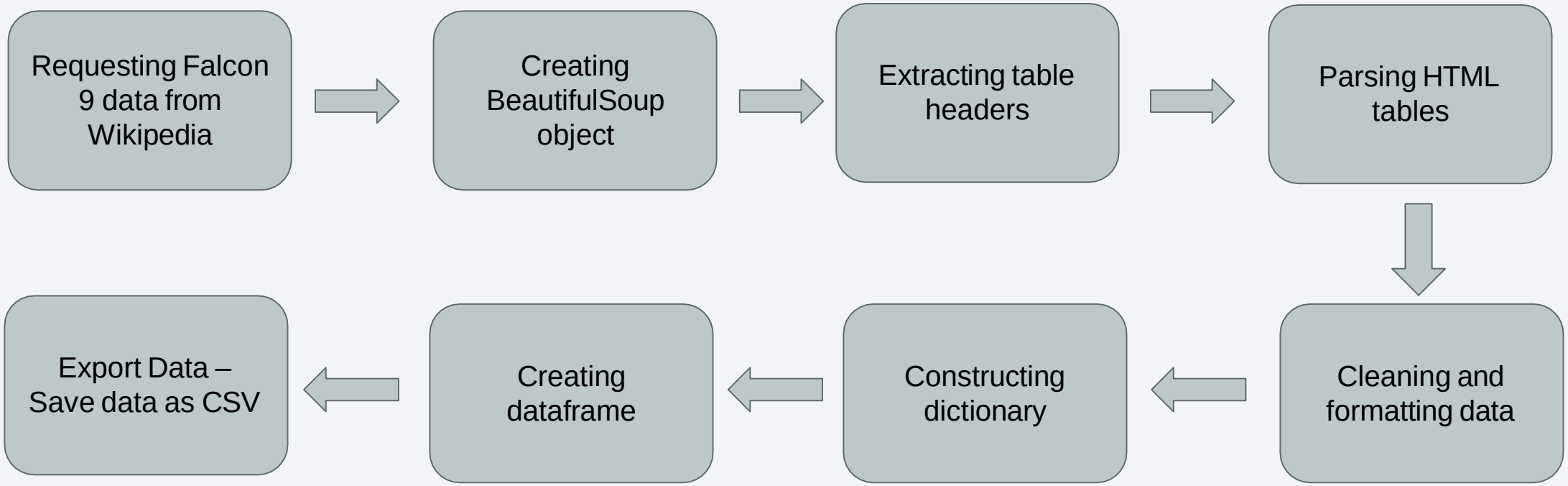
Data from SpaceX was gathered from two sources: one from the **SpaceX API**, which provided past launch details, and the other by **scraping a Wikipedia snapshot**, which contained historical launch records for further analysis.

Data Collection – SpaceX API



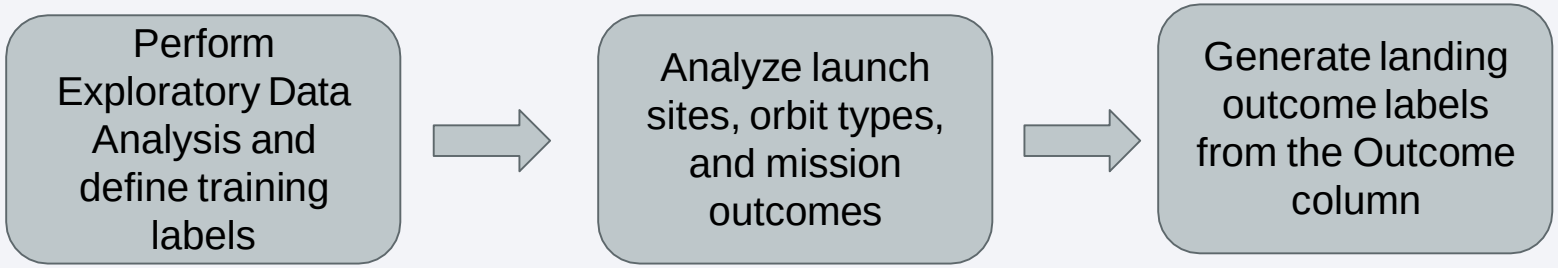
GitHub Url:
https://github.com/Ameya32/Applied-Data-Science-Capstone-Project/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



GitHub Url:
https://github.com/Ameya32/Applied-Data-Science-Capstone-Project/blob/main/2_jupyter-labs-webscraping.ipynb

Data Wrangling



Initially, Exploratory Data Analysis (EDA) was performed to understand the dataset. The number of launches per site, occurrences of each orbit, and mission outcomes were analyzed. Booster landings were classified as successful or failed based on landing methods (Ocean, RTLS, ASDS). Finally, these outcomes were converted into training labels, with "1" for success and "0" for failure.

GitHub Url:
https://github.com/Ameya32/Applied-Data-Science-Capstone-Project/blob/main/3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

To explore the data, various charts were plotted, such as Flight Number vs. Payload Mass, Launch Site vs. Payload Mass, Orbit Type vs. Success Rate, and yearly success trends. Scatter plots helped identify relationships between variables, which could be useful for machine learning.

Bar charts were used to compare discrete categories like launch sites and orbit types, while line charts captured trends over time, showing changes in success rates across different years.

GitHub Url:
https://github.com/Ameya32/Applied-Data-Science-Capstone-Project/blob/main/5_edadataviz.ipynb

EDA with SQL

Performed SQL queries:

Display the names of the unique launch sites in the space mission

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

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

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

List the date when the first successful landing outcome in ground pad was achieved

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

List the total number of successful and failure mission outcomes

- List the names of the booster versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- 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

GitHub Url:
https://github.com/Ameya32/Applied-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

Launch Site Visualization:

- **Mapping Launch Sites**: Markers were added to indicate the locations of all launch sites, including NASA Johnson Space Center, using their latitude and longitude coordinates. These markers help visualize their proximity to the equator and coastal regions, which are important factors for rocket launches.
- **Visualizing Launch Outcomes**: Colored markers were used to distinguish between successful and failed launches—green for success and red for failure. These markers were grouped into clusters to easily identify which launch sites have higher success rates.
- **Analyzing Proximity to Key Locations**: Colored lines were added to illustrate distances between launch sites and nearby landmarks. For example, the KSC LC-39A site was connected to its closest railway, highway, coastline, and city to analyze accessibility and logistical advantages.

GitHub Url:
https://github.com/Ameya32/Applied-Data-Science-Capstone-Project/blob/main/6_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Interactive Data Visualization:

Launch Site Selection: A dropdown list was implemented to allow users to select a specific launch site for analysis.

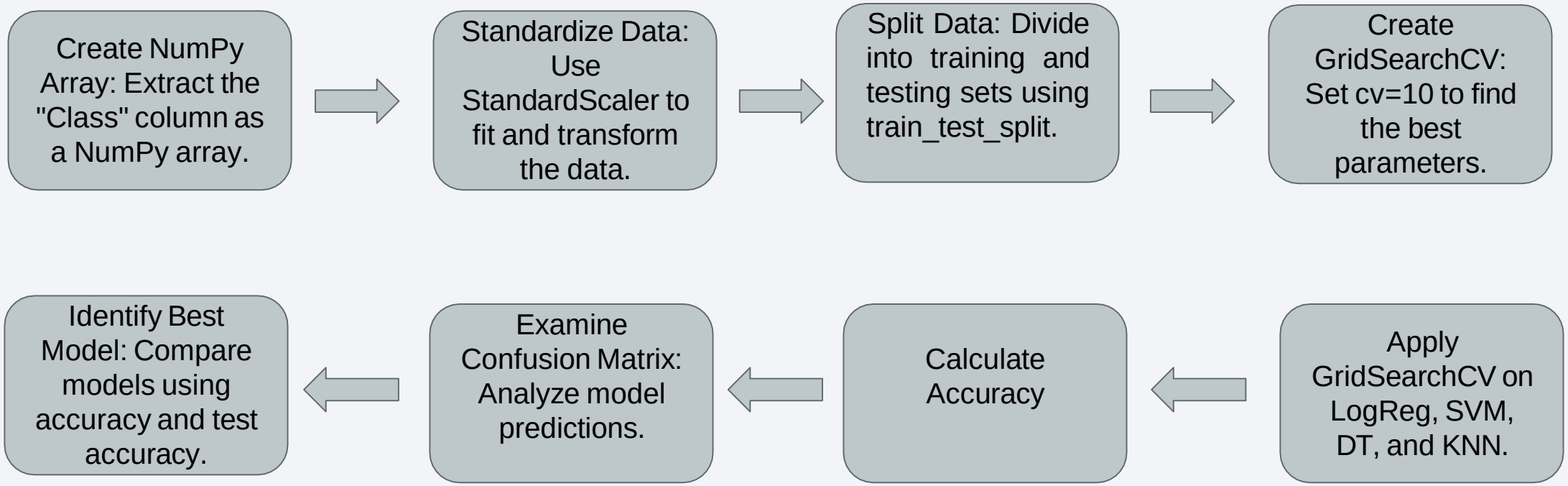
Success Rate Visualization: A pie chart was added to display the total successful launches for all sites. When a specific site is selected, it shows the success vs. failure distribution for that site.

Payload Mass Filtering: A slider was introduced to enable users to filter data based on payload mass range, allowing for more focused analysis.

P**ayload vs. Success Correlation**: A scatter plot was created to visualize the relationship between payload mass and success rate across different booster versions, helping to identify trends in launch performance.

GitHub Url:
https://github.com/Ameya32/Applied-Data-Science-Capstone-Project/blob/main/spacex_dash_app.py

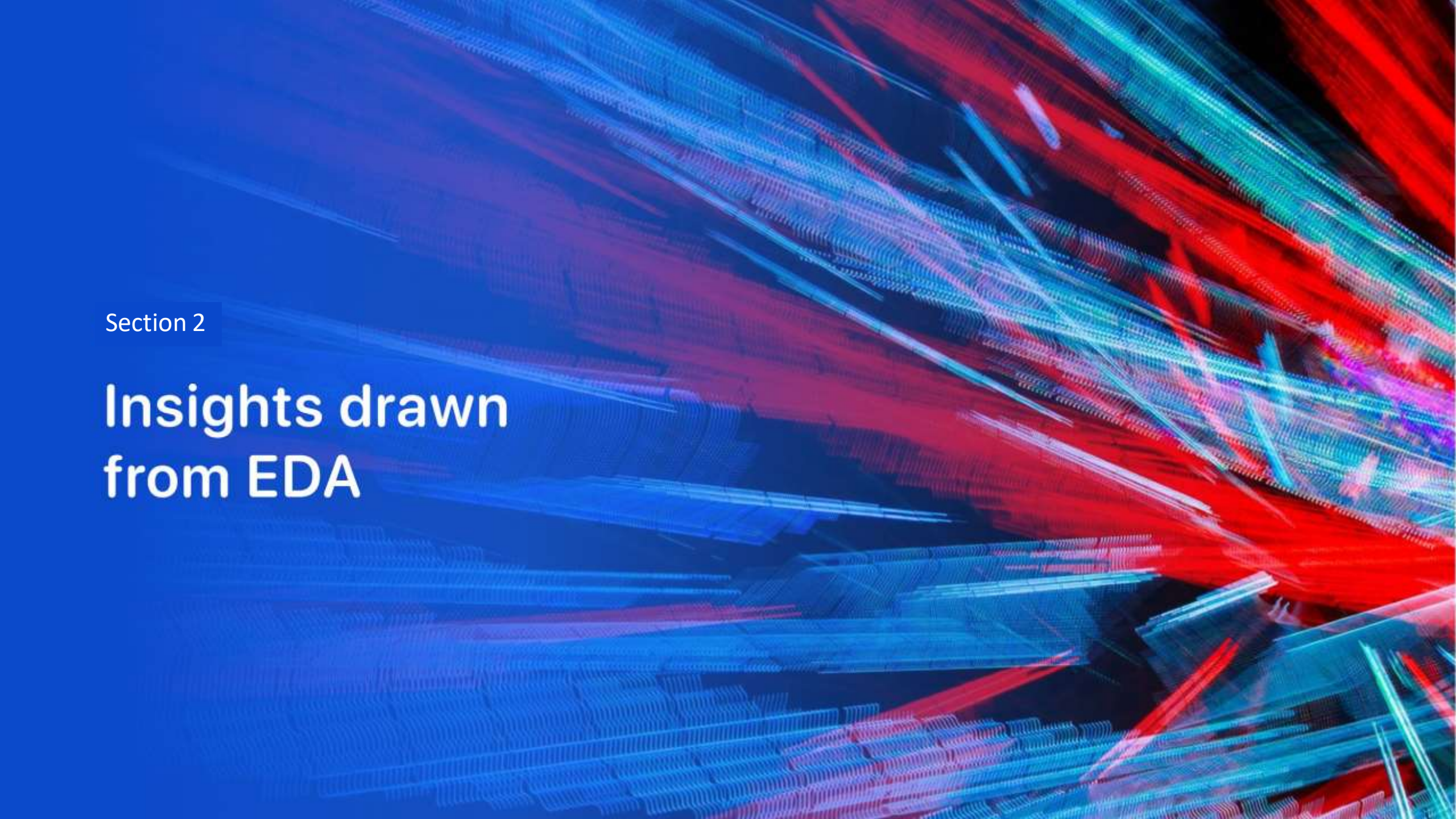
Predictive Analysis (Classification)



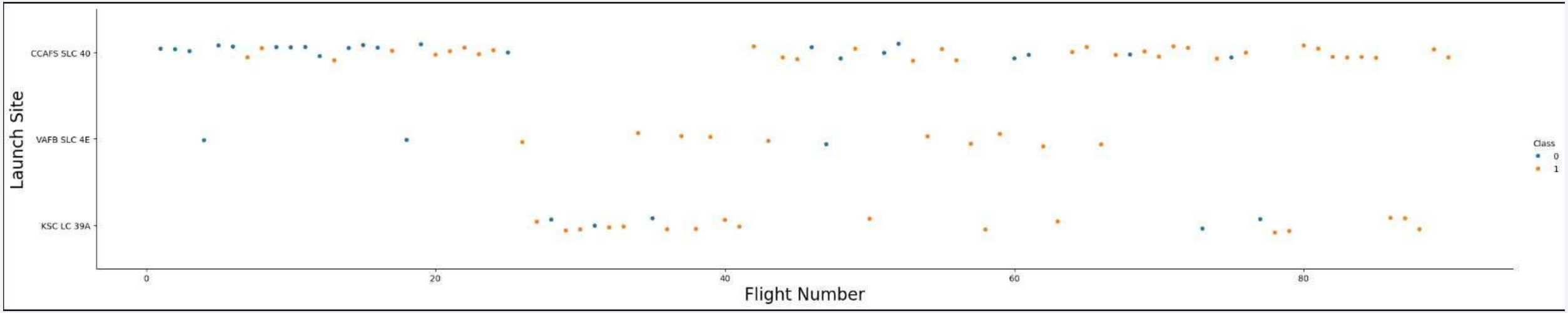
GitHub Url:
https://github.com/Ameya32/Applied-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
  - SpaceX has 4 launch sites, with early launches done for NASA and itself.
  - The first successful landing occurred in 2015, five years after the first launch.
  - Most Falcon 9 boosters successfully landed on drone ships, especially with payloads above the 2,928 kg average.
  - Landing success rates improved over time, with almost 100% mission success, except for two failed landings in 2015 (F9 v1.1 B1012 & B1015).
- Interactive analytics demo in screenshots
  - Launch sites are strategically located near the sea for safety and have strong logistic infrastructure, with most launches occurring on the East Coast.
- Predictive analysis results
  - The Decision Tree model achieved the highest accuracy (0.8875), while LogReg, SVM, and KNN performed similarly around 0.848.


Section 2
Insights drawn
from EDA

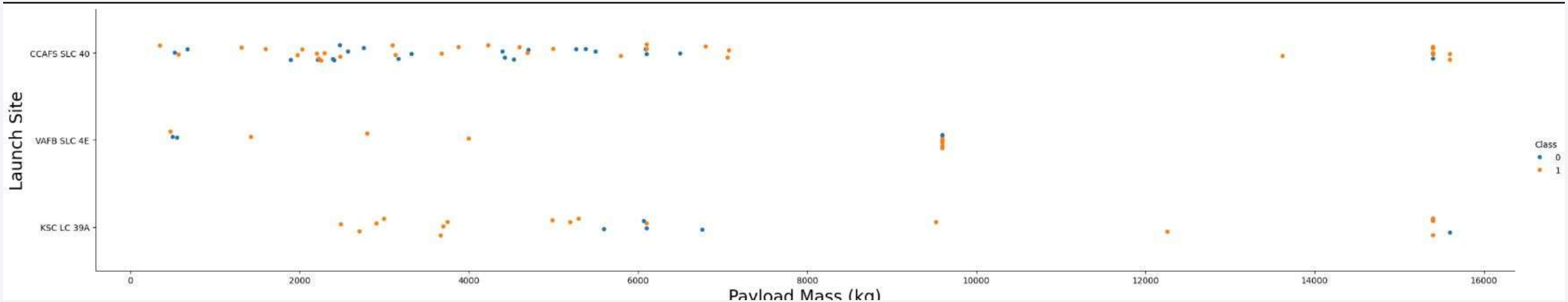
Flight Number vs. Launch Site



Explanation:

- Early flights failed, while recent ones succeeded.
- CCAFS SLC 40 handled nearly half of all launches.
- VAFB SLC 4E and KSC LC 39A had higher success rates.
- Success rates improved with newer launches.

Payload vs. Launch Site



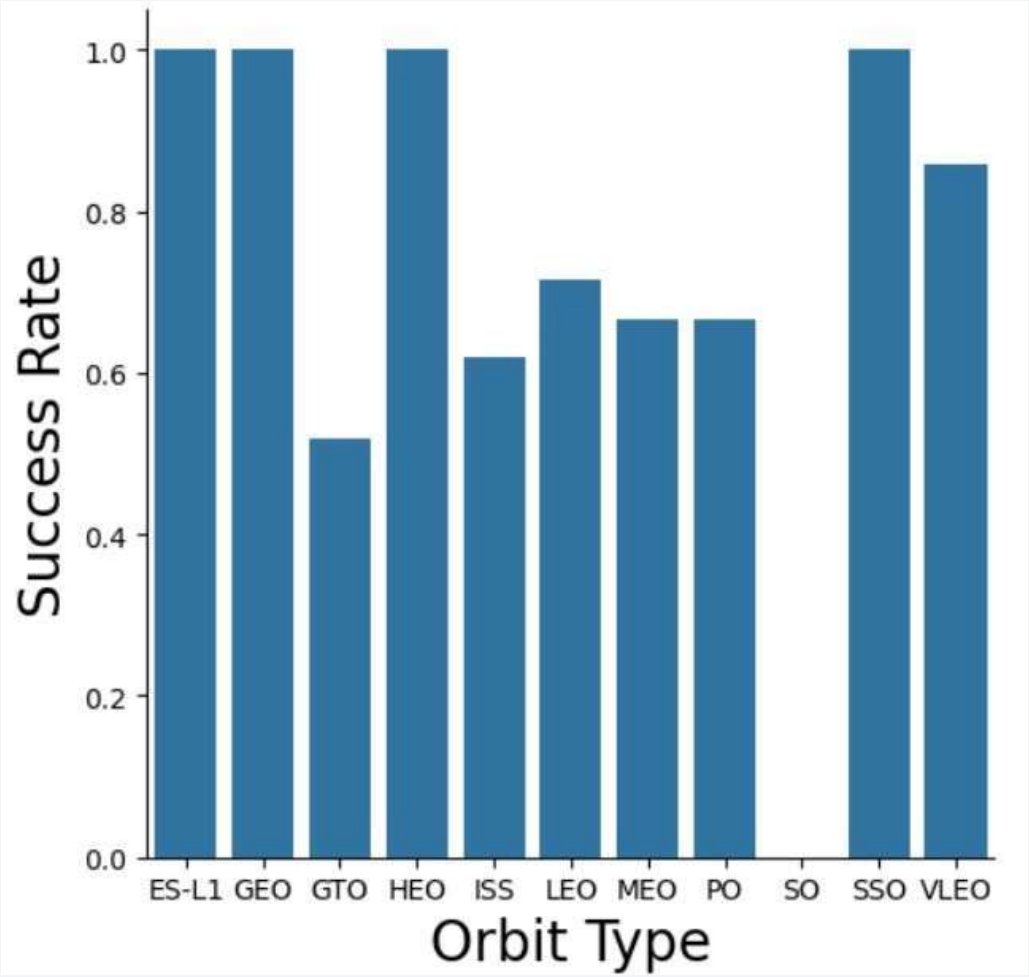
Explanation:

- Payloads over 9,000 kg have excellent success rate;
- Payloads over 12,000 kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites

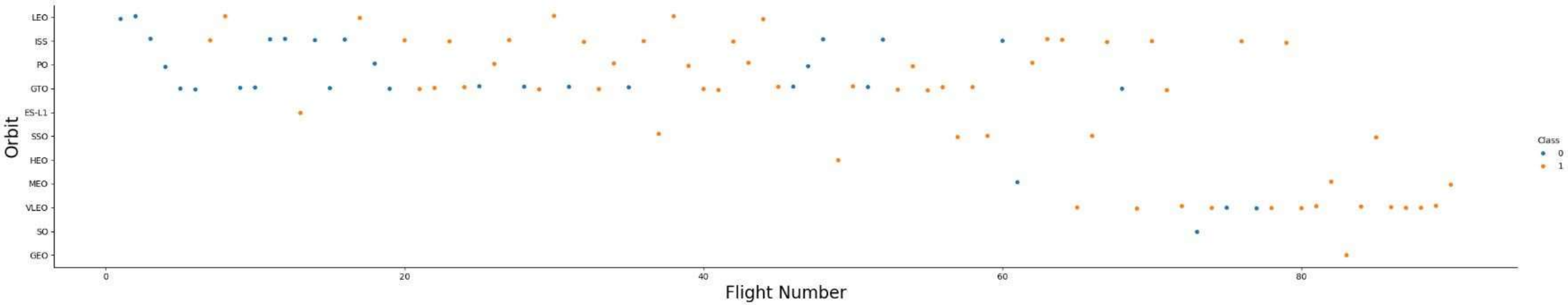
Success Rate vs. Orbit Type

Explanation:

- 100% success rate: ES-L1, GEO, HEO, SSO.
- 0% success rate: SO.
- 50% to 85% success rate: GTO, ISS, LEO, MEO, PO.



Flight Number vs. Orbit Type



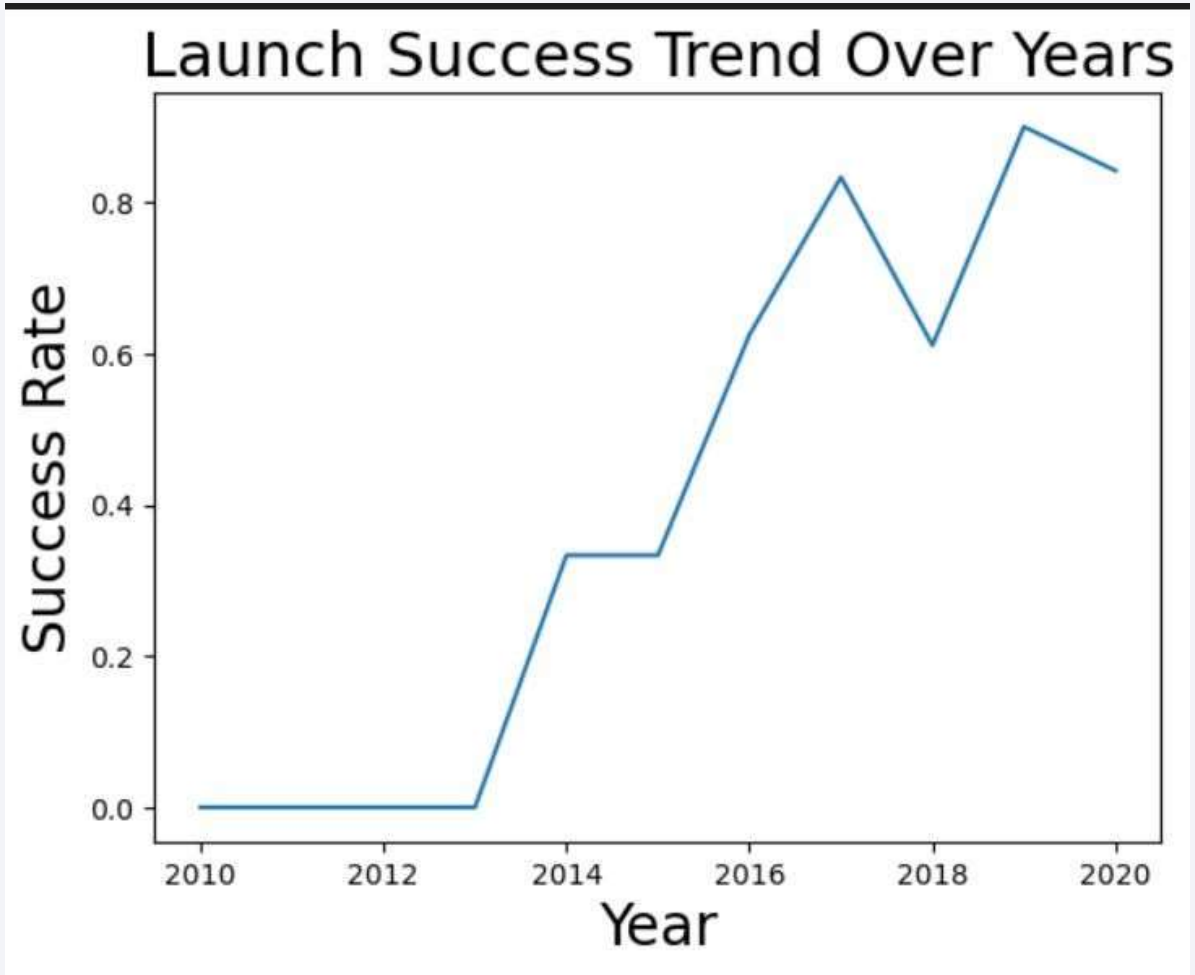
Explanation:

- Apparently, the success rate has improved over time for all orbits.
- VLEO orbit seems to be a new business opportunity due to its recent increase in frequency.

Launch Success Yearly Trend

Explanation:

- The success rate steadily increased from 2013 to 2020.



All Launch Site Names

According to data set, there are four launch sites

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

Query : select distinct launch_site from SPACEXTBL;

Launch Site Names Begin with 'CCA'

Top 5 records where launch sites begin 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

Query : select * from SPACEXTBL where launch_site like 'CCA%' limit 5;

Total Payload Mass

Total payload carried by boosters

total_payload_mass
45596

Query : select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1:

average_payload_mass
2928.4

Query : select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version = 'F9 v1.1';

First Successful Ground Landing Date

First successful landing outcome on ground pad:

first_successful_landing
2015-12-22

Query : select min(date) as first_successful_landing from SPACEXTBL where LANDING_OUTCOME = 'Success (ground pad)';

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Query : select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;

Total Number of Successful and Failure Mission Outcomes

Number of successful and failure mission outcomes:

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

Query : select mission_outcome, count(*) as total_number from SPACEXTBL 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

Query : select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);

2015 Launch Records

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Query :SELECT SUBSTR(Date, 6, 2) AS Month, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Failure (drone ship)'
AND SUBSTR(Date, 1, 4) = '2015';

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

Ranking of all landing outcomes between the date 2010-06-04 and 2017 03-20:

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

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

Section 3

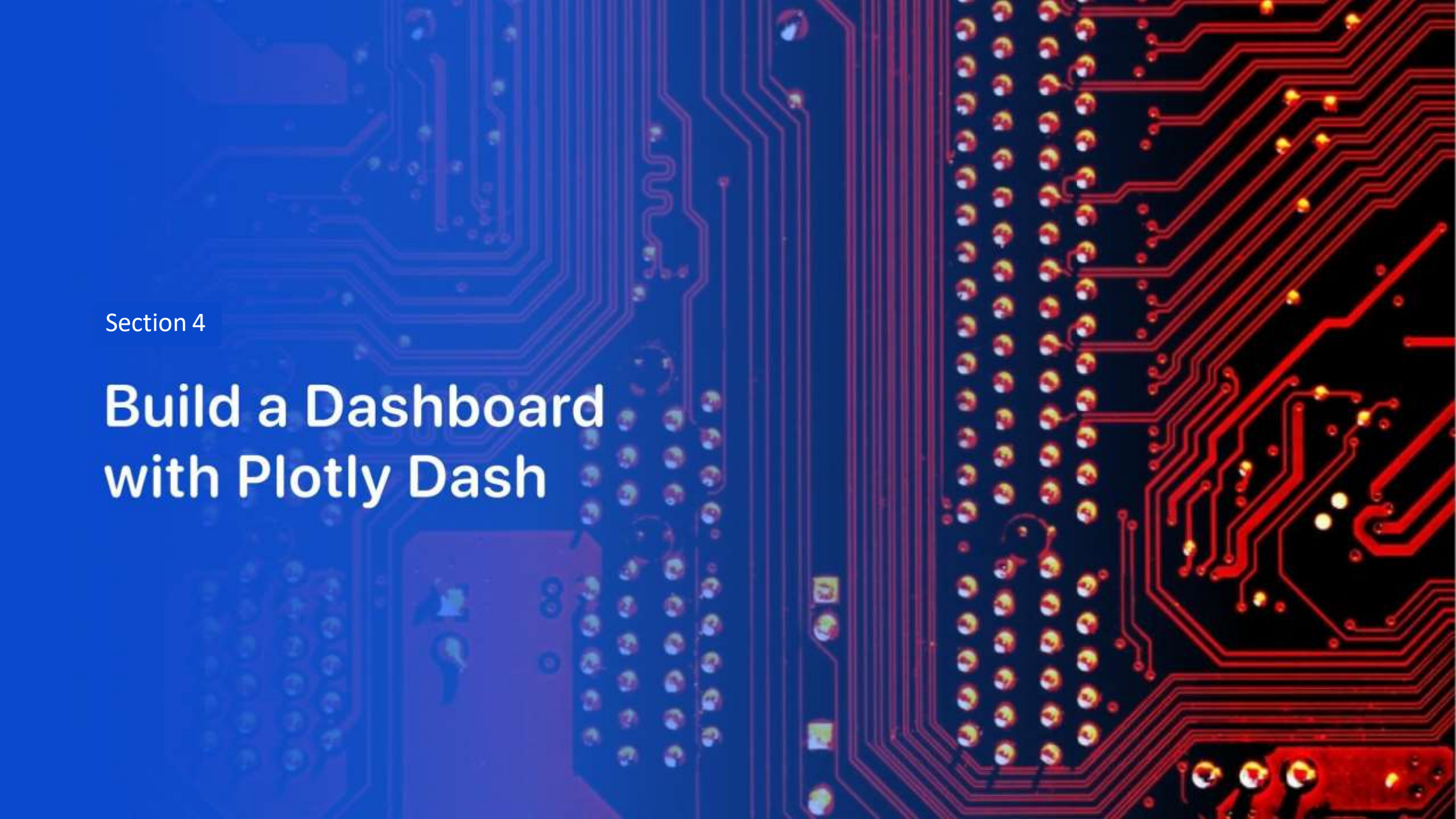
Launch Sites Proximities Analysis

Folium Map Screenshot 1

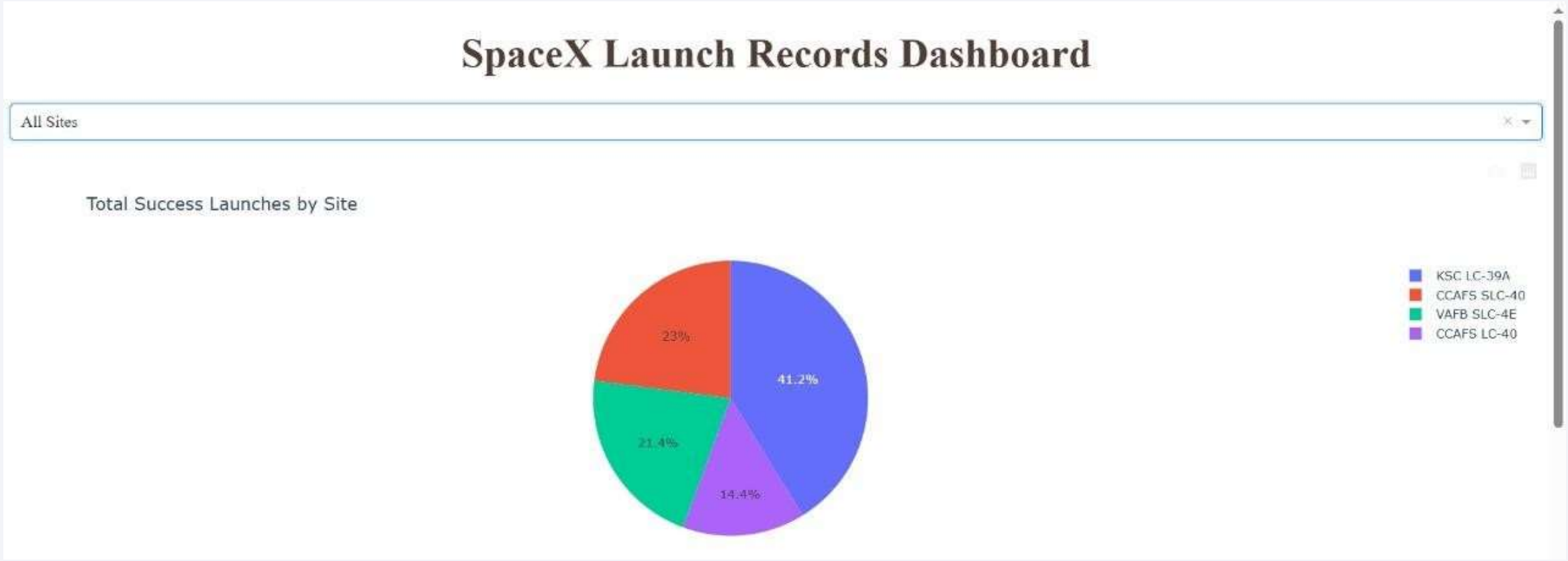
All launch sites are near the coast to minimize risks, ensuring debris or explosions occur over the ocean away from populated areas.

Folium Map Screenshot 2

Folium Map Screenshot 3

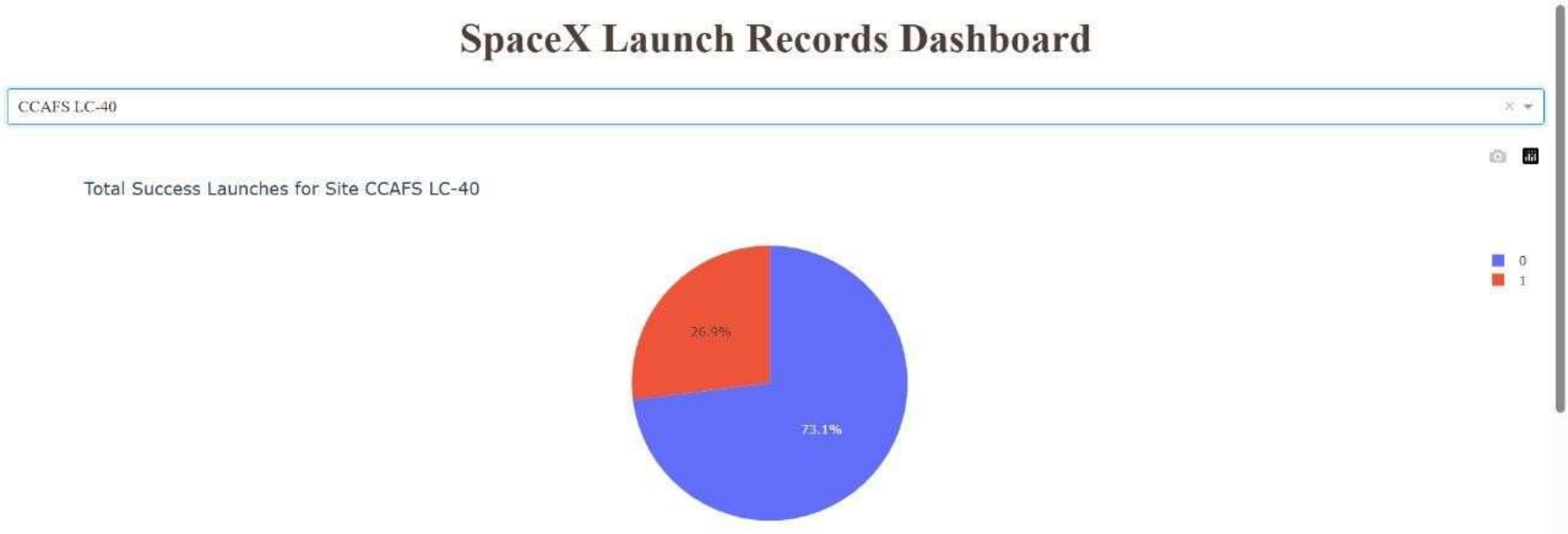

Section 4
Build a Dashboard with Plotly Dash

Dashboard Screenshot 1



The pie chart displays the total number of successful launches across all launch sites.

Dashboard Screenshot 2



The pie chart represents the total number of successful launches from the CCAFS LC-40 launch site

Dashboard Screenshot 3

The scatter plot shows the correlation between payload mass and launch success across different booster versions.

Section 5

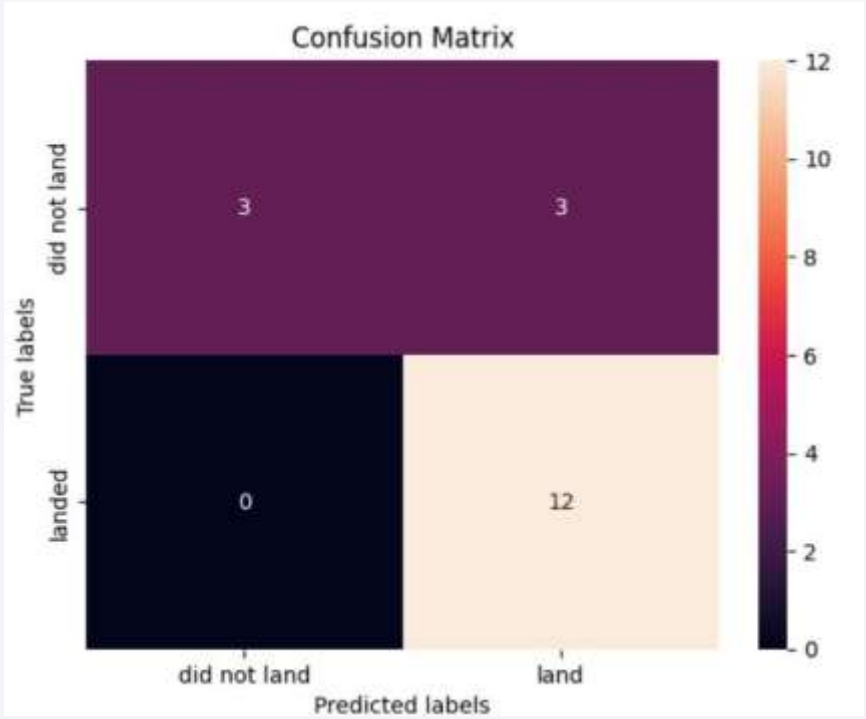
Predictive Analysis (Classification)

Classification Accuracy

```
Model           Accuracy        TestAccuracy
LogReg          0.84643         0.83333
SVM             0.84821         0.83333
Tree            0.8875          0.83333
KNN             0.84821         0.83333
Tuned hyperparameters (best parameters) for LogReg: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
Accuracy for LogReg: 0.84643
Tuned hyperparameters (best parameters) for SVM: {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
Accuracy for SVM: 0.84821
Tuned hyperparameters (best parameters) for Tree: {'criterion': 'gini', 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_spli
t': 5, 'splitter': 'random'}
Accuracy for Tree: 0.8875
Tuned hyperparameters (best parameters) for KNN: {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
Accuracy for KNN: 0.84821
```

The classification models achieved high accuracy, with the Decision Tree model performing the best at 88.75%, followed by SVM and KNN at 84.82%, and Logistic Regression at 84.64%. Despite slight variations, all models showed consistent test accuracy of 83.33%, indicating good generalization. Hyperparameter tuning improved model performance, with key parameters such as regularization strength (C) for Logistic Regression, kernel and gamma for SVM, depth and splitting criteria for Decision Tree, and neighbor count for KNN playing crucial roles. The results suggest that tree-based models may be more effective for this classification task.

Confusion Matrix



The confusion matrix demonstrates the model's accuracy by highlighting a higher number of correct predictions (true positives and true negatives) compared to incorrect ones (false positives and false negatives).

Conclusions

- The **Decision Tree model may perform best** for this dataset
- Lower payload mass launches tend to have higher success rates.
- Launch sites are mostly near the Equator and close to the coast for safety and logistics.
- Launch success rates have improved over the years.
- KSC LC-39A has the highest success rate among all launch sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.

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!