IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marcus Lew
22.08.2023

Outline

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

Executive Summary

This reports seeks to identify the factors for a successful rocket landing.

Collection: used SpaceX REST API and web scraping techniques (BeautifulSoup)

Data wrangling: to create success/fail outcome variable
Exploratory data analysis:visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend

Analysis: Used SQL, statistics provided: total payload, total no. of successful and failed outcomes, payload range for successful launches

Data exploration: launch site success rates and proximity to geographical markers

Visualisations:launch sites with highest success and range of successful payloads

Models built: logistic regression, support vector machine (SVM), K -nearest neighbor (KNN), decision tree

Exploratory Data Analysis:
Launch success has improved over time
KSC LC-39A has the highest success rate among landing sites
Orbits ES -L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics:
Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:
All models performed similarly on the test set. The decision tree slightly outperformed

Introduction

SpaceX, a prominent figure in the field of space exploration, is dedicated to achieving the goal of making space travel accessible to a broader population. Among its notable achievements are successfully delivering spacecraft to the international space station, initiating the deployment of a constellation of satellites designed to offer global internet coverage, and orchestrating crewed missions into space. The driving force behind SpaceX's ability to accomplish these feats lies in its innovative approach to cost-effectiveness, as evidenced by the relatively economical nature of its rocket launches, priced at $62 million per launch. This remarkable affordability can be attributed to the ingenious strategy of recycling the initial stage of its Falcon 9 rocket. In contrast, competing space service providers, lacking the capability for first stage reuse, incur considerably higher expenses, with costs exceeding $165 million per launch.

The determining factor for the launch cost hinges on the successful recovery and reuse of the first rocket stage. This pivotal aspect can be gauged by evaluating whether the first stage achieves a controlled landing. To facilitate this evaluation, a combination of publicly available data and advanced machine learning models can be employed to predict the likelihood of SpaceX, or a rival company, effectively reusing the initial rocket stage.

Section 1

Methodology

Methodology

- Process the data - including data filtration, treatment of missing values, and implementation of one-hot encoding - to ready it for analysis and modeling.
- Investigate the data through exploratory data analysis (EDA) employing SQL and data visualization methods.
- Employ Folium and Plotly Dash for data visualization purposes.
- Construct prediction models for landing outcomes using classification algorithms. Refine and assess these models to identify the optimal model and parameter settings.

Data Collection - SpaceXAPI

https://github.com/ML3782/IBMFinalDataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Procedure:

- Initiate a data request from the SpaceX API to obtain rocket launch data.
- Decode the received response using the .json() function and then transform it into a dataframe using .json_normalize().
- Retrieve launch-specific details from the SpaceX API through custom functions designed for this purpose.
- Transform the acquired data into a dictionary structure.
- Convert the dictionary data into a dataframe format.
- Apply filtering to the dataframe, retaining only entries corresponding to Falcon 9 launches.
- For any instances of missing values in the Payload Mass column, substitute them with the computed mean value.
- Save the processed data by exporting it to a CSV file.

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- https://github.com/ML3782/IBMFinalDataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb

Procedure:

• Initiate a data request from Wikipedia to retrieve Falcon 9 launch data.
• Construct a BeautifulSoup object by parsing the HTML response obtained.
• Extract the column names by analyzing the header of the HTML table.
• Retrieve the relevant data by parsing the HTML tables on the page.
• Organize the collected data into a dictionary format.
• Transform the dictionary-based data into a dataframe structure.
• Store the processed data by exporting it to a CSV file.

Data Wrangling

Steps:

• Conduct Exploratory Data Analysis (EDA) to understand the data and identify data labels.

• Perform calculations for the following metrics:

Number of launches for each launch site

Number and occurrence of orbits

Number and occurrence of mission outcomes per orbit type

Introduce a binary landing outcome column as the dependent variable.

Save the processed data into a CSV file.

Landing was not always successful

“True Ocean” shows mission outcome was a successful landing to a region of the ocean

"False Ocean" signifies an unsuccessful landing in a designated ocean region.

"True RTLS" indicates a successful landing on a ground pad for the mission.

"False RTLS" corresponds to an unsuccessful landing on a ground pad.

"True ASDS" indicates a successful landing on a drone ship for the mission.

"False ASDS" signifies an unsuccessful landing on a drone ship.

Mission outcomes are converted to 1 for successful landings and 0 for unsuccessful landings.

https://github.com/ML3782/IBMFinalDataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts:

- Plot of Flight Number against Payload
- Plot of Flight Number against Launch Site
- Plot of Payload Mass (kg) against Launch Site
- Plot of Payload Mass (kg) against Orbit Type

https://github.com/ML3782/IBMFinalDataScienceCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

Exploratory Data Analysis (EDA) with Visualization:

Analysis:

- Utilize scatter plots to visually explore relationships between variables. If a discernible relationship is present, these variables could prove valuable for machine learning purposes.
- Utilize bar charts to compare discrete categories. Bar charts effectively illustrate relationships among different categories and their corresponding measured values.

EDA with SQL

https://github.com/ML3782/IBMFinalDataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- Queries:
- Unique launch site names
- 5 records where the launch site starts with 'CCA'
- Total payload mass carried by boosters launched by NASA under the CRS program
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of the first successful landing on a ground pad
- Names of boosters that achieved successful landings on a drone ship, with payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions that carried the maximum payload
- Failed landing outcomes on drone ships, along with their booster version and launch site, for the months within the year 2015
- Count of landing outcomes between June 4, 2010, and March 20, 2017, ordered in descending order.

Build an Interactive Map with Folium

Markers Indicating Launch Sites:

• Placed a blue circle marker on the coordinates of NASA Johnson Space Center, accompanied by a popup label indicating its name, utilizing its latitude and longitude coordinates.

• Included red circle markers on the coordinates of all launch sites, each featuring a popup label showcasing its name, utilizing its latitude and longitude coordinates.

Map with Folium:

Colored Markers of Launch Outcomes:

• Incorporated colored markers on the map to represent successful launches (green) and unsuccessful launches (red) at each corresponding launch site, offering a visual indication of the success rates of these sites.

Distances Between a Launch Site to Proximities:

• Introduced colored lines on the map to depict the distances from launch site CCAFS SLC-40 to its closest points of interest, including the nearest coastline, railway, highway, and city. This feature provides insight into the site's proximity to various important landmarks.

https://github.com/ML3782/IBMFinalDataScienceCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Dropdown List for Launch Sites:

• Implemented a dropdown list enabling users to select either all available launch sites or a specific launch site of interest.

Dashboard Using Plotly Dash:

Slider for Payload Mass Range:

• Integrated a slider component into the dashboard, granting users the ability to define a desired payload mass range.

Pie Chart Displaying Successful Launches:

• Enabled users to view the distribution of successful and unsuccessful launches through a pie chart, representing these outcomes as a percentage of the total number of launches.

Scatter Chart Depicting Payload Mass vs. Success Rate by Booster Version:

• Provided users with a scatter chart that showcases the relationship between payload mass and launch success rate for different booster versions, allowing for visual analysis of this correlation.

https://github.com/ML3782/IBMFinalDataScienceCapstone/blob/main/Plotlygraph.py

Predictive Analysis (Classification)

Charts:

- Generate a NumPy array based on the Class column.
- Perform data standardization using StandardScaler. Fit the scaler to the data and then transform it.
- Divide the data into training and testing sets using the train_test_split function.
- Set up a GridSearchCV object with cross-validation of 10 folds to optimize parameters.
- Apply GridSearchCV using different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), and K-Nearest Neighbor (KNeighborsClassifier()).
- Evaluate the accuracy of all models on the test data using the .score() method.
- Analyze the confusion matrix for each model.
- Determine the optimal model based on Accuracy metrics.

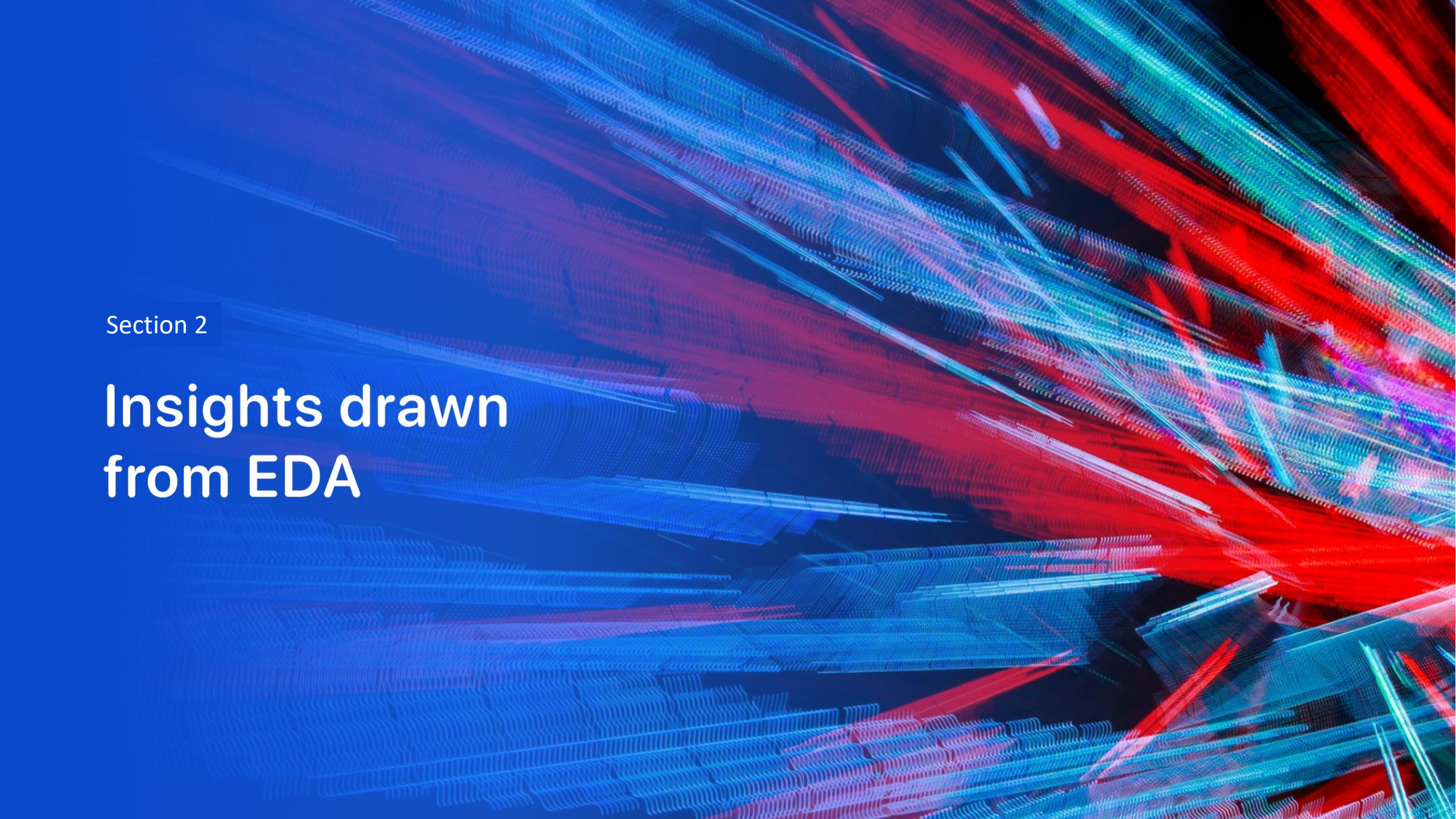
https://github.com/ML3782/IBMFinalDataScienceCapstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

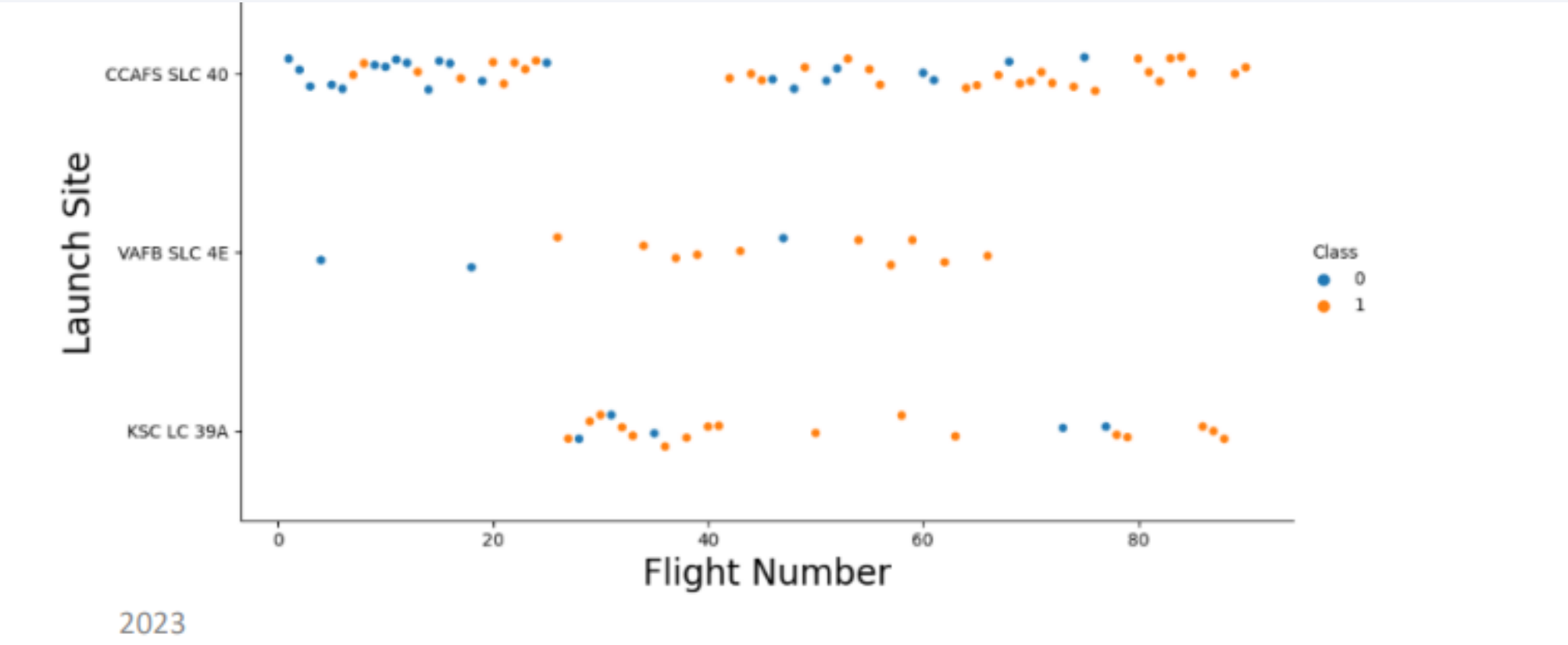
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results• Notable improvement in launch success rates observed over time.
- • KSC LC-39A stands out with the highest success rate among all landing sites.
- • Orbits such as ES-L1, GEO, HEO, and SSO exhibit a perfect 100% success rate.

- Results Summary - Visual Analytics:
- • Most launch sites are strategically situated near the equator, and all are located in close proximity to coastlines.
- • Launch sites are adequately distanced from urban areas, highways, and railways to mitigate potential damage in the event of a launch failure. Nonetheless, they remain sufficiently accessible for logistical support.

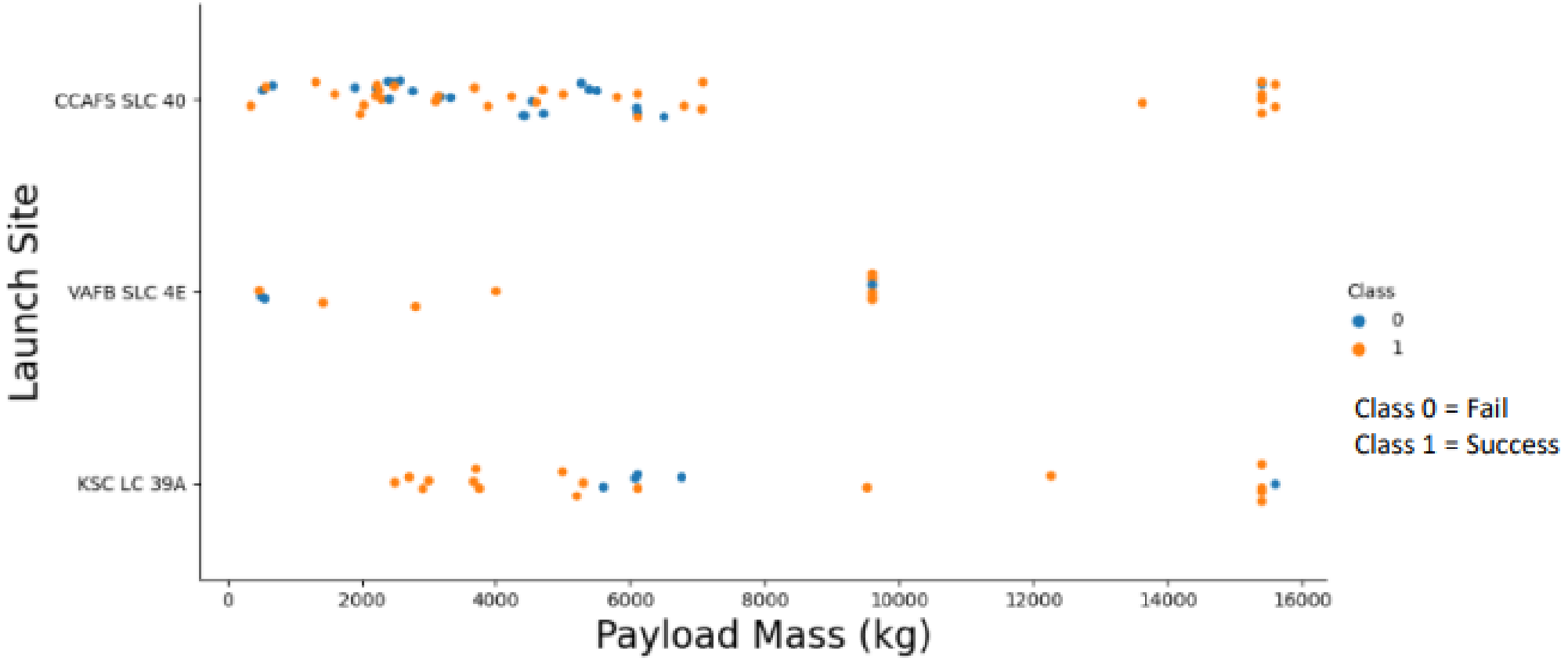
- Predictive Analytics:
- • The Decision Tree model emerges as the most effective predictive model for the dataset, offering the highest accuracy in making predictions.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

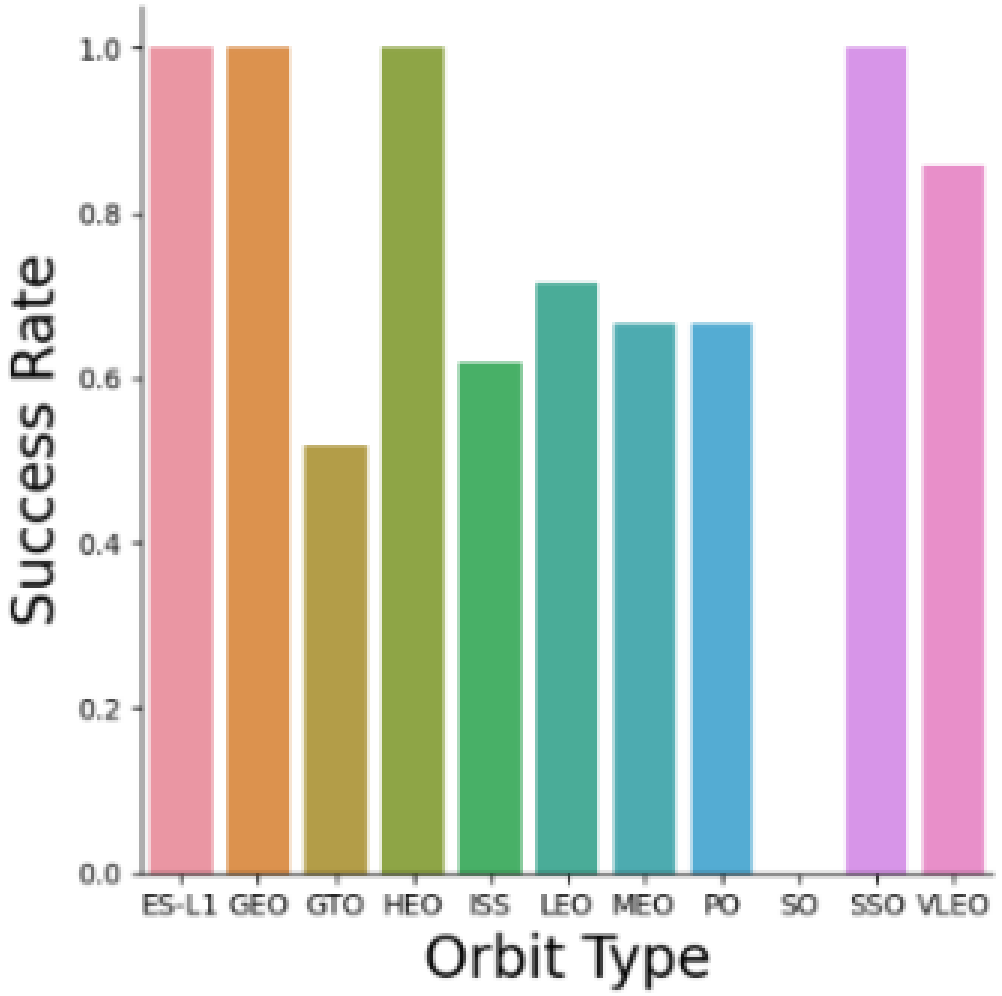


Payload vs. Launch Site

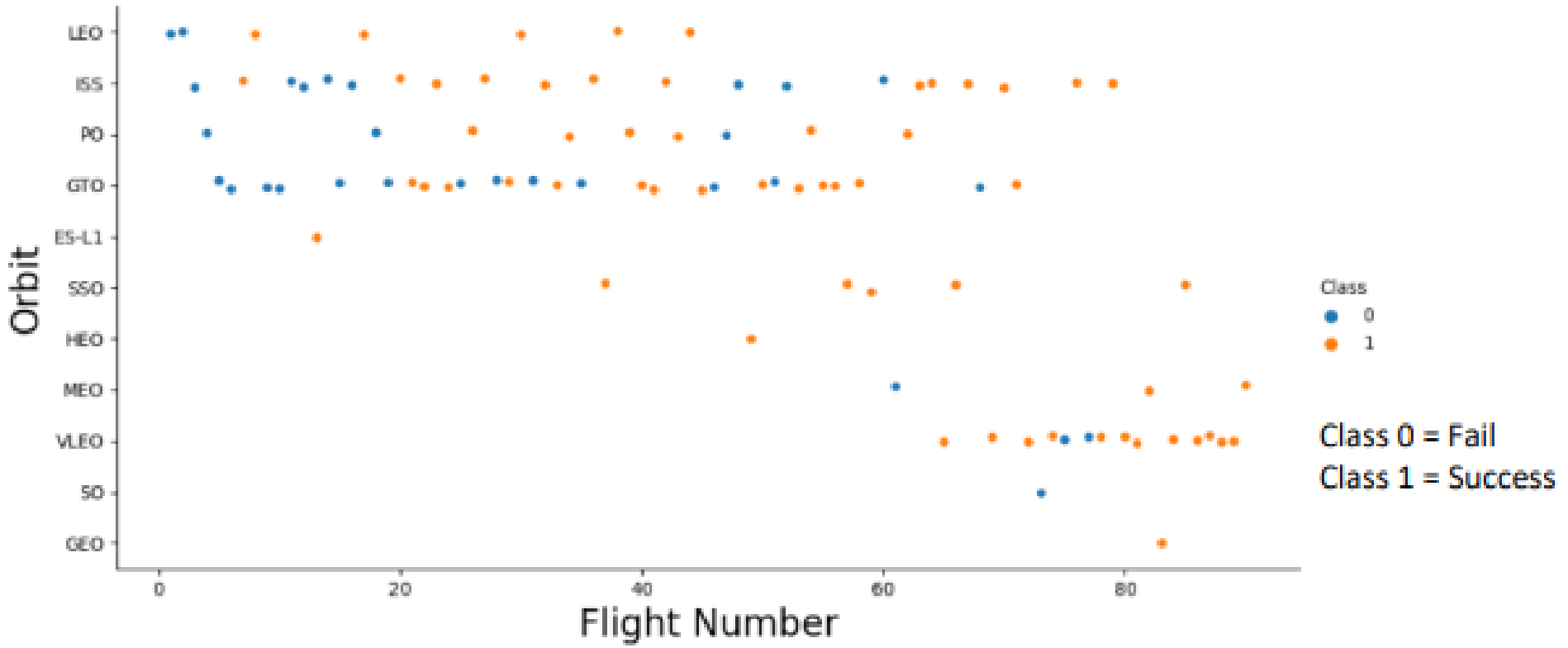


2023

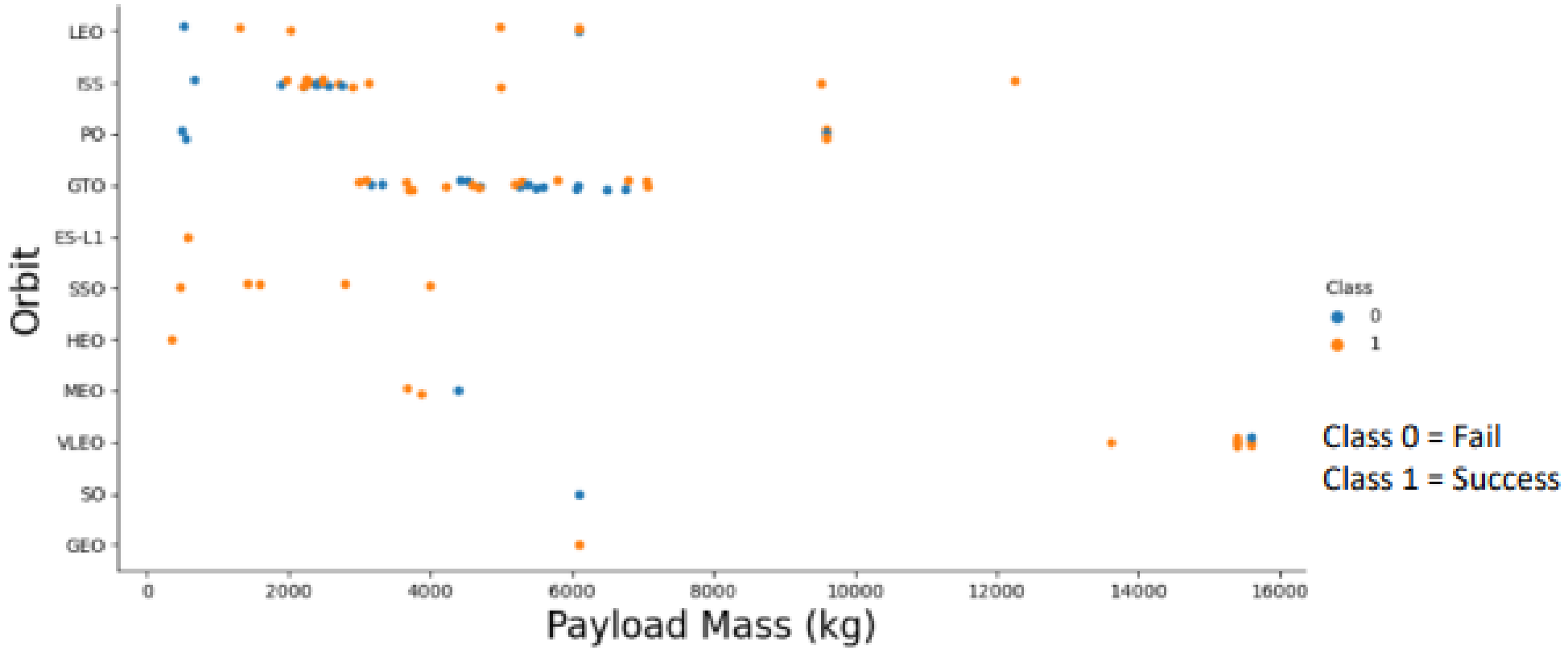
Success Rate vs. Orbit Type



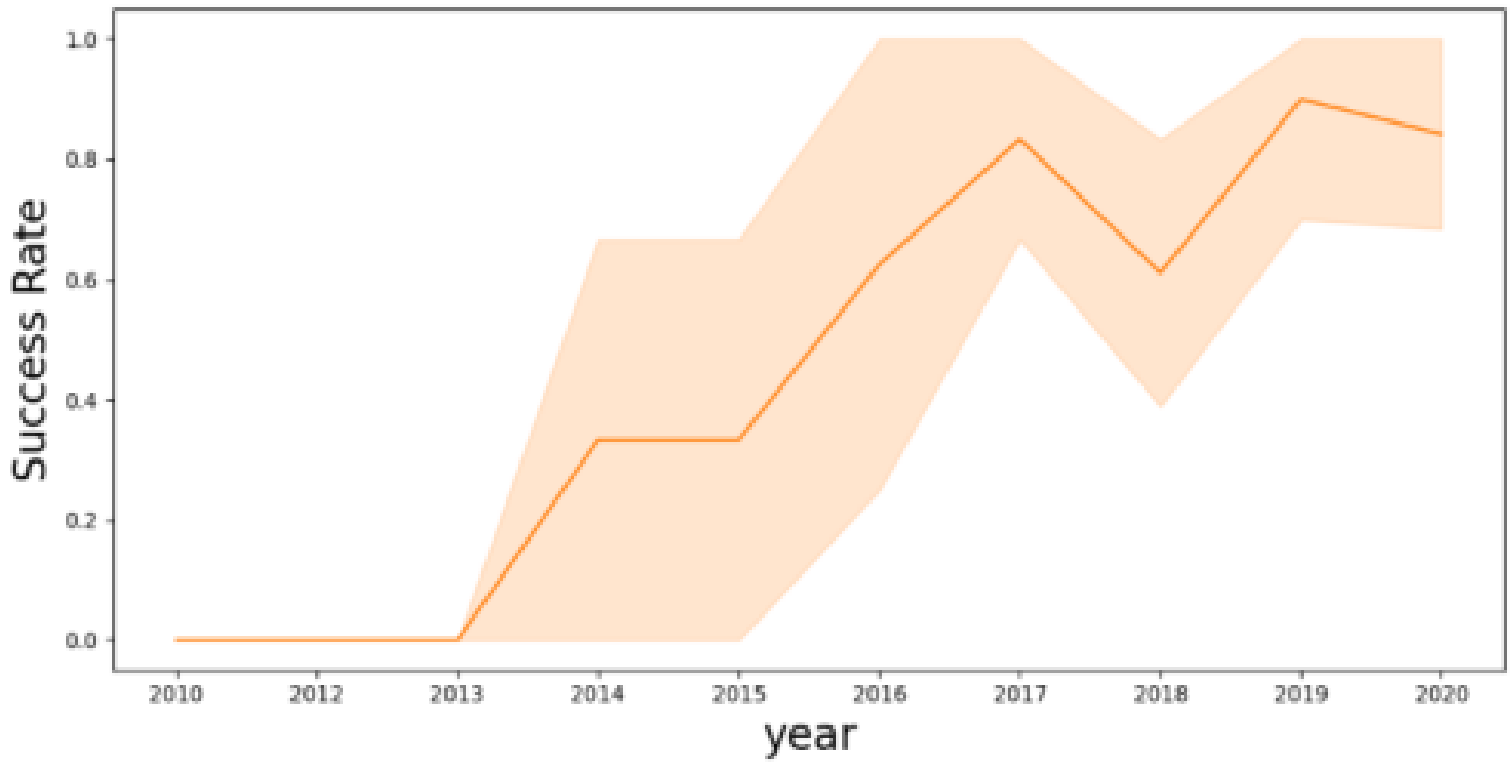
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- Find the names of the unique launch sites
- CCAFS LC-40 • CCAFS SLC-40 • KSC LC-39A • VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Find 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

1
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

1
2928

v1.1

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

1
2015-12-22

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9
   sqlite:///my_data1.db
Done.
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

-
-

Mission_Outcome	total_number
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 th d mass
- Present your query result with a short explanation h

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

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Rank the count of landing outcc
  (ground pad)) between the date
  descending order

- Present your query result with a

done.

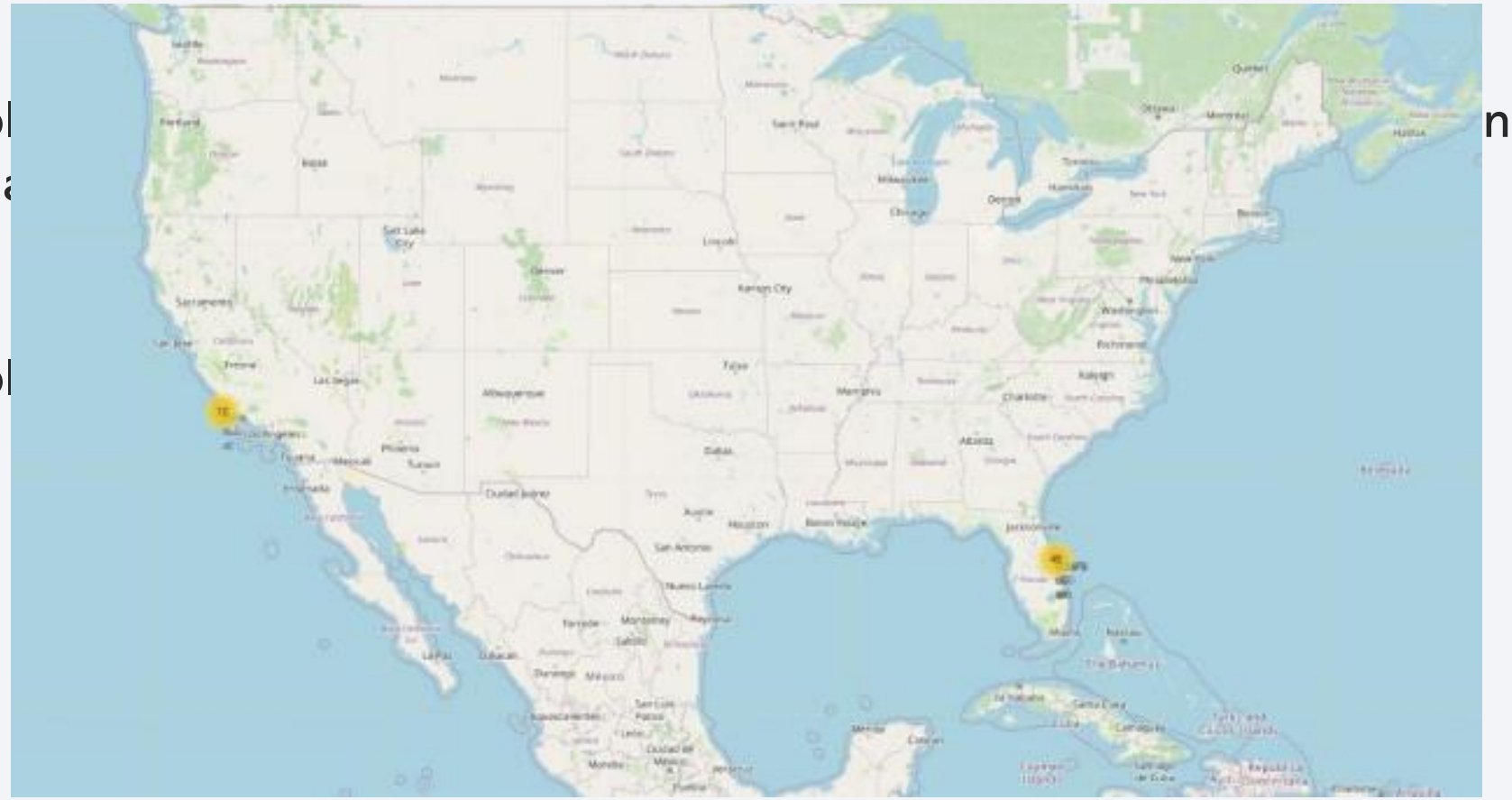
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Folium Map of USA and launch sites

- Replace <Folium map screenshot 1> title with an appropriate title

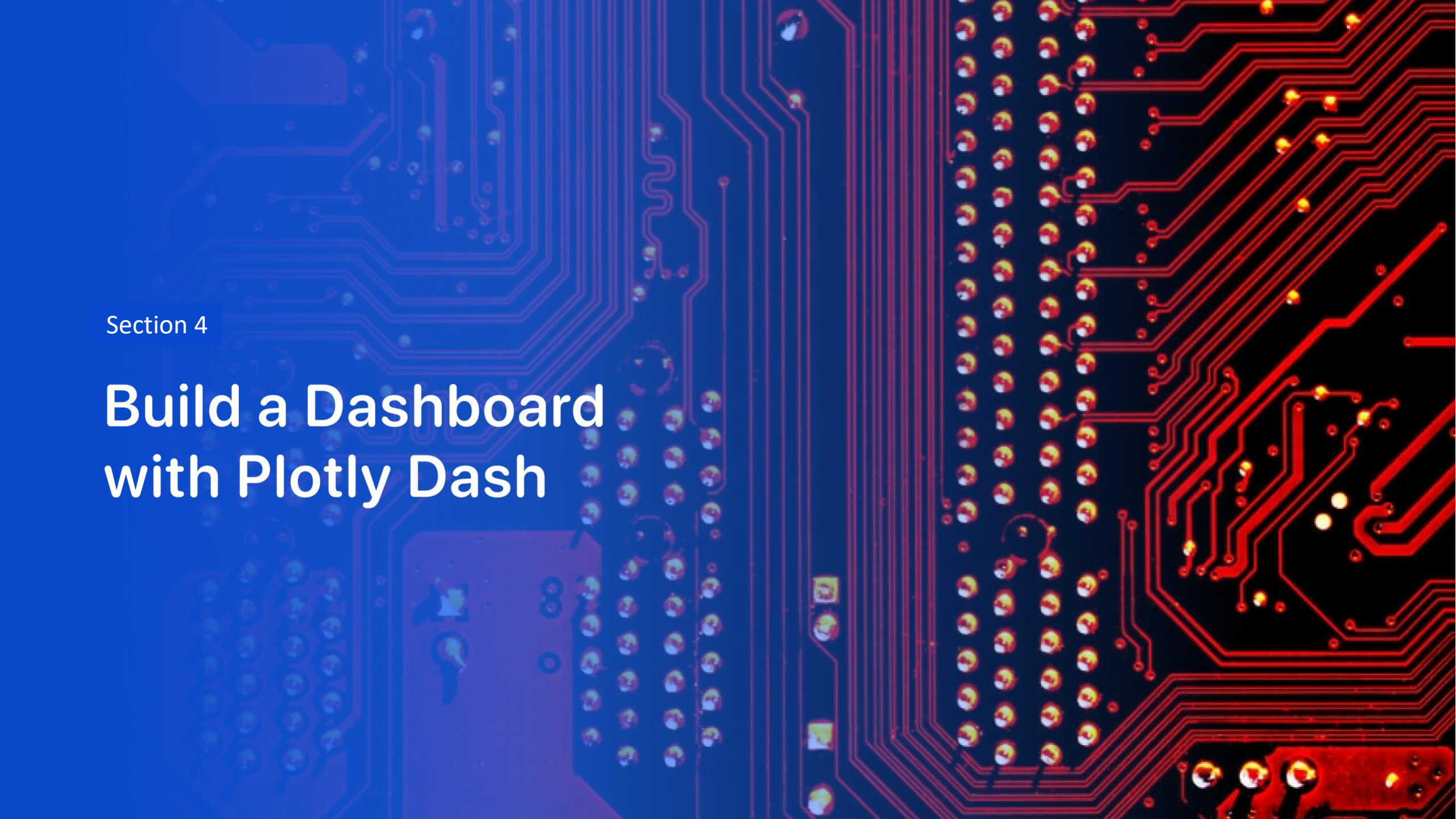


- Expl nclude all la
- Expl

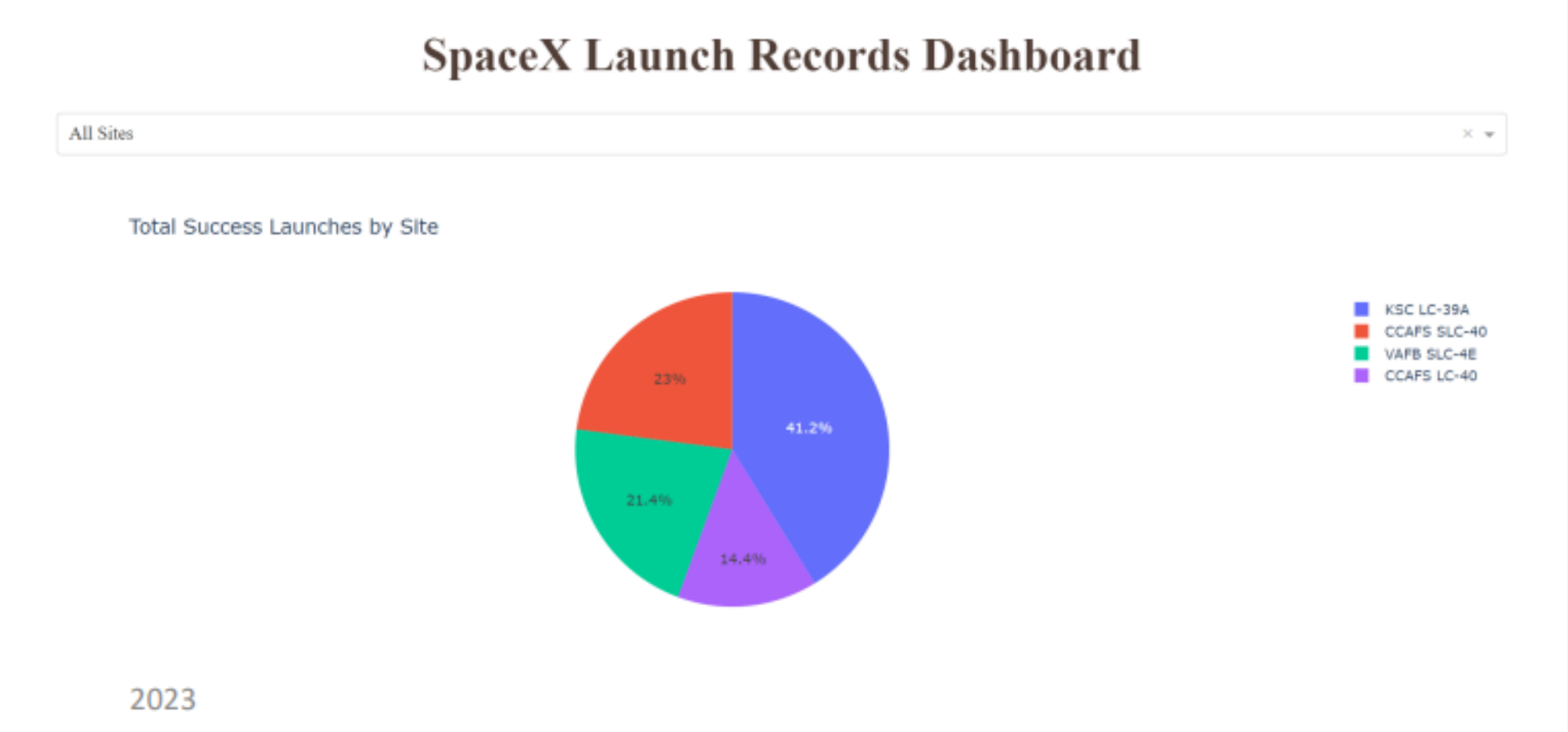
Folium Map of USA and launch sites

<Folium Map Screenshot 3>

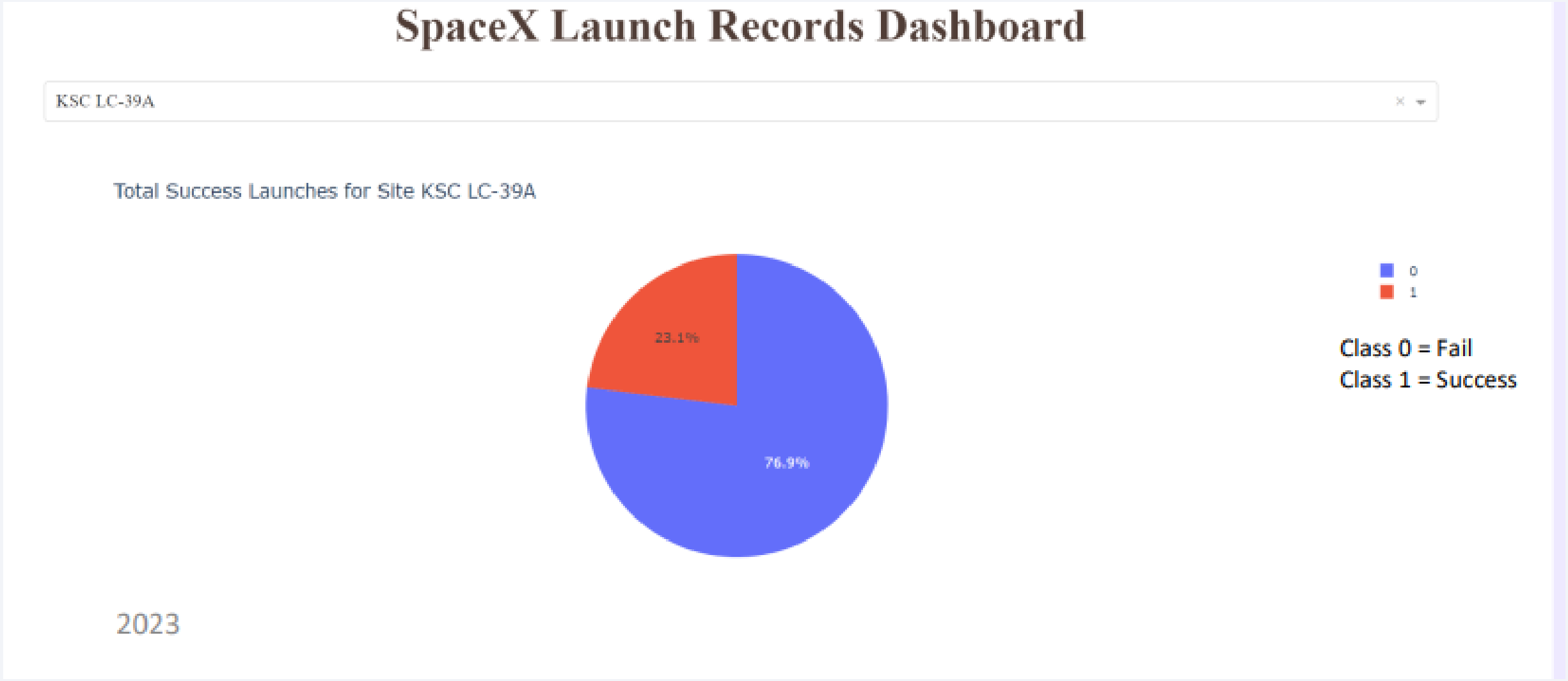
- Replace <Folium map screenshot 3> title with an appropriate title


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



<Dashboard Screenshot 2>



<Dashboard Screenshot 3>

Section 5

Predictive Analysis (Classification)

Classification Accuracy

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

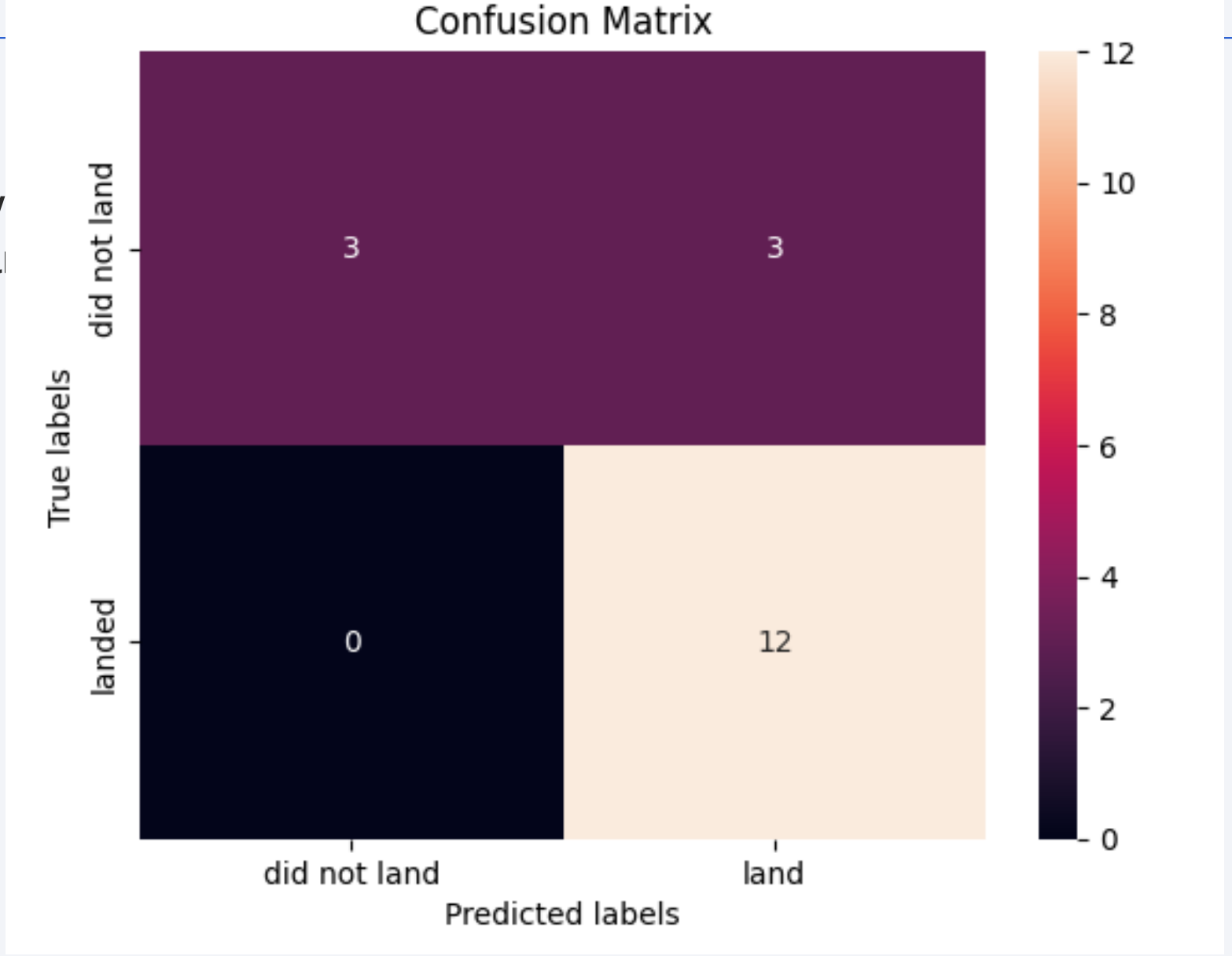
- Find which model has the highest classification accuracy

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitter': 'random'}
accuracy : 0.9
```

Confusion Matrix

- Show
  expla



Conclusions

- Model Performance: The models exhibited relatively similar performance on the test set, with the decision tree model showing a slight edge in performance.

Equator: A majority of the launch sites are strategically located near the equator. This positioning capitalizes on the Earth's rotational speed, providing a natural boost that reduces the need for extra fuel and boosters, resulting in cost savings.

- Coast:• All launch sites are situated in close proximity to coastlines, facilitating access to waterways for launch activities.

- Launch Success: Over time, there is an observable trend of increasing launch success rates.

- KSC LC-39A: The KSC LC-39A launch site stands out with the highest success rate among all launch sites. For launches with payloads less than 5,500 kg, this site maintains a perfect 100% success rate.

- Orbits: Orbits such as ES-L1, GEO, HEO, and SSO boast a flawless 100% success rate, further underscoring their reliability.

- Payload Mass: Across all launch sites, there exists a correlation: as the payload mass (measured in kg) increases, so does the likelihood of a successful launch.

Thank you!