IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

George Hoholis
6/20/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of Methodologies
  - Data Collection and Preparation:
    - Data Source: We used SpaceX launch data which included information on launch sites, payload masses, and launch outcomes.
    - Data Cleaning: The data was cleaned and preprocessed to handle missing values, ensure consistency, and prepare it for analysis and visualization.
  - Geospatial Analysis and Mapping:
    - Folium for Mapping: We utilized the Folium library to create interactive maps showing the locations of SpaceX launch sites.
    - Circle and Marker Visualization: Each launch site was marked with a circle and a label to visualize the site's geographical distribution.
  - Interactive Dashboard Development:
    - Dash Framework: A Dash application was created to allow interactive exploration of the SpaceX launch data.
  - Predictive Modeling:
    - Classification Models: Several classification models were built to predict the success of SpaceX launches
    - Model Training and Testing: The models were trained and tested on historical launch data to evaluate their performance.
    - Accuracy Evaluation: The accuracy of each model was calculated to compare their predictive performance.

Executive Summary

- Summary of Results
  - Geospatial Insights: The interactive map provided a clear visualization of the geographical distribution of SpaceX launch sites, allowing for easy identification of site locations and their respective launch success rates.
  - Launch Site Performance: The pie charts indicated that while some launch sites had higher success rates than others, the overall success rate of SpaceX launches was significant.
  - Payload vs. Launch Success:
    - The scatter plot analysis revealed a correlation between payload mass and launch success, suggesting that payload weight might influence the outcome of launches.
    - This relationship helps in optimizing payload capacities to increase the likelihood of successful launches.
  - Predictive Model Performance:
    - The classification models showed varying levels of accuracy in predicting launch success:
    - The Decision Tree model exhibited the highest accuracy among the models tested.

Introduction

- The commercial space age is ushering in an era where space travel becomes affordable for everyone. Companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX are leading the charge. A key factor in SpaceX's cost efficiency is the reusability of the Falcon 9 rocket's first stage, significantly reducing launch costs.
- Therefore if we can determine if the Falcon 9's first stage will land successfully, this will be key to reducing launch costs.

Section 1

Methodology

Methodology

Executive Summary

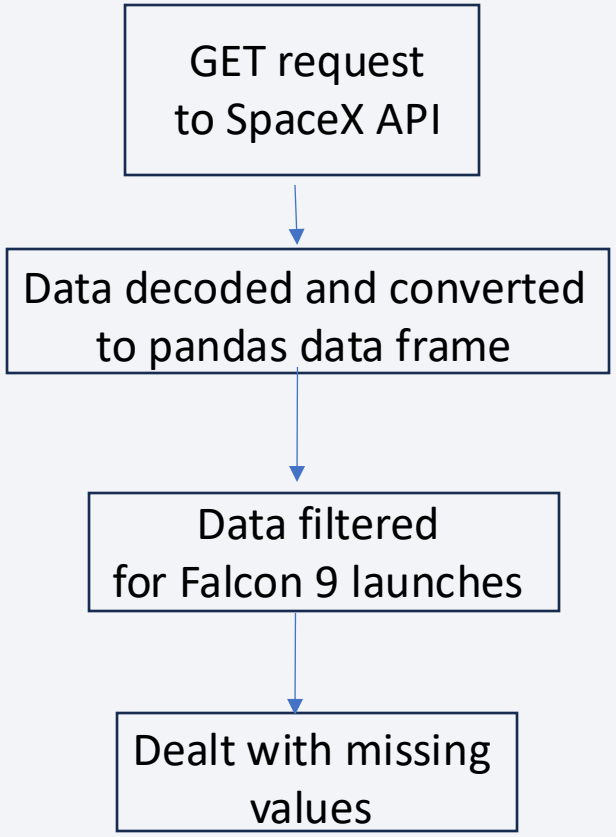
- Data collection methodology:
  - Space X launch data was collected from the Space X API
- Perform data wrangling
  - The data was filtered to include only Falcon 9 launches and then missing values in the payload mass were replaced with the mean.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- To accurately predict the landing success of SpaceX's Falcon 9 first stage and determine launch costs, we followed a structured data collection process:
  - Web Scraping: Extracted historical launch data from SpaceX's official website and other space-related databases using web scraping techniques.
  - API Integration: Utilized SpaceX API to obtain detailed information on each launch, including launch dates, payloads, launch sites, and outcomes.
  - Data Cleaning and Preprocessing: Filtered, cleaned, and standardized the data to ensure consistency and accuracy.

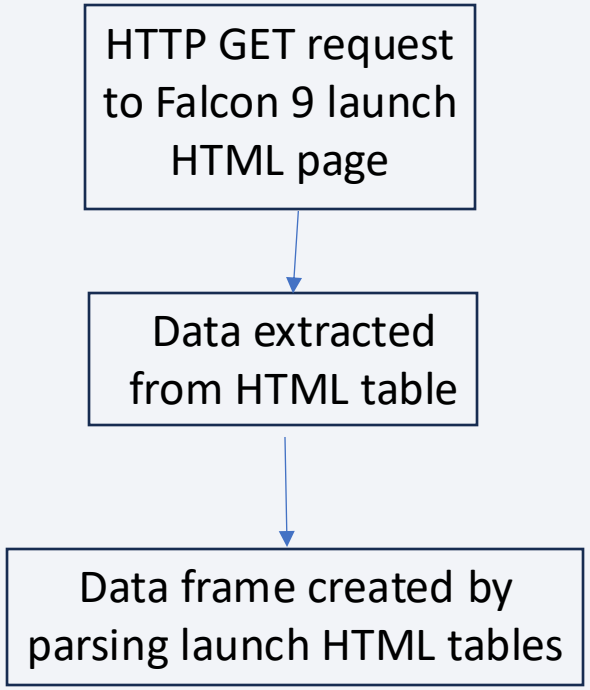
Data Collection – SpaceX API

- The launch data was obtained from the SpaceX API using a GET request.
- This was then decoded using .json() and and turned into a pandas data frame with .json_normalize().
- It was then filtered to include only Falcon 9 launches.
- Lastly missing values in the payload mass were replaced with the mean of the column.
- The GitHub URL of the completed SpaceX API calls notebook is https://github.com/GHoholis/AppliedDataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api_complete.ipynb



Data Collection - Scraping

- Data for the Falcon 9 launches was scraped from a wikipedia page titled "list of Falcon 9 and Falcon Heavy launches" using and HTTP GET method.
- The data was extracted from the launch HTML table and parsed to create a data frame.
- The GitHub URL of the completed web scraping notebook is https://github.com/GHoholis/AppliedDataScienceCapstone/blob/main/jupyter-labs-webscraping_complete.ipynb



Data Wrangling

- To get an initial understanding of the data the launches for each site, the occurrences of each orbit and the occurrences of the mission outcomes were calculated.
- Then a landing outcome label was created from the mission outcomes to be used for later classification.
- The GitHub URL of the completed data wrangling notebooks is https://github.com/GHoholis/AppliedDataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling_complete.ipynb

EDA with Data Visualization

- A variety of charts were plotted during the EDA these include scatter plots, bar charts and line graphs.
- A scatter plot was used to visualize the relationship between launch site and payload mass, flight number and payload mass, and flight number and launch site.
- A bar chart was used to visualize the relationship between success rate and orbit type.
- A line graph was used to visualize the relationship between the average success rate of launches and the year they were launched.
- The GitHub URL of the completed EDA with data visualization notebook is https://github.com/GHoholis/AppliedDataScienceCapstone/blob/main/jupyter-labs-eda-dataviz_complete.ipynb

EDA with SQL

- Several SQL queries were performed to get an understanding of the data, these include:
  - The names of the unique launch sites.
  - The total payload mass carried by booster launched by NASA.
  - The average payload mass carried by booster F9 v1.1.
  - The date the first successful landing in a ground pad.
  - The total number of successful and failure mission outcomes.
  - The booster versions which carried the maximum payload mass.
- The GitHub URL of the completed EDA with SQL notebook is https://github.com/GHoholis/AppliedDataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite_complete.ipynb

Build an Interactive Map with Folium

- Circles and markers were added for each of the four launch sites, with labels containing their names.
- Then marker clusters were added for the launches colored green/red to indicate success/failure, respectively.
- The GitHub URL of the completed interactive map with Folium map is https://github.com/GHoholis/AppliedDataScienceCapstone/blob/main/lab_jupyter_launch_site_location_complete.ipynb

Build a Dashboard with Plotly Dash

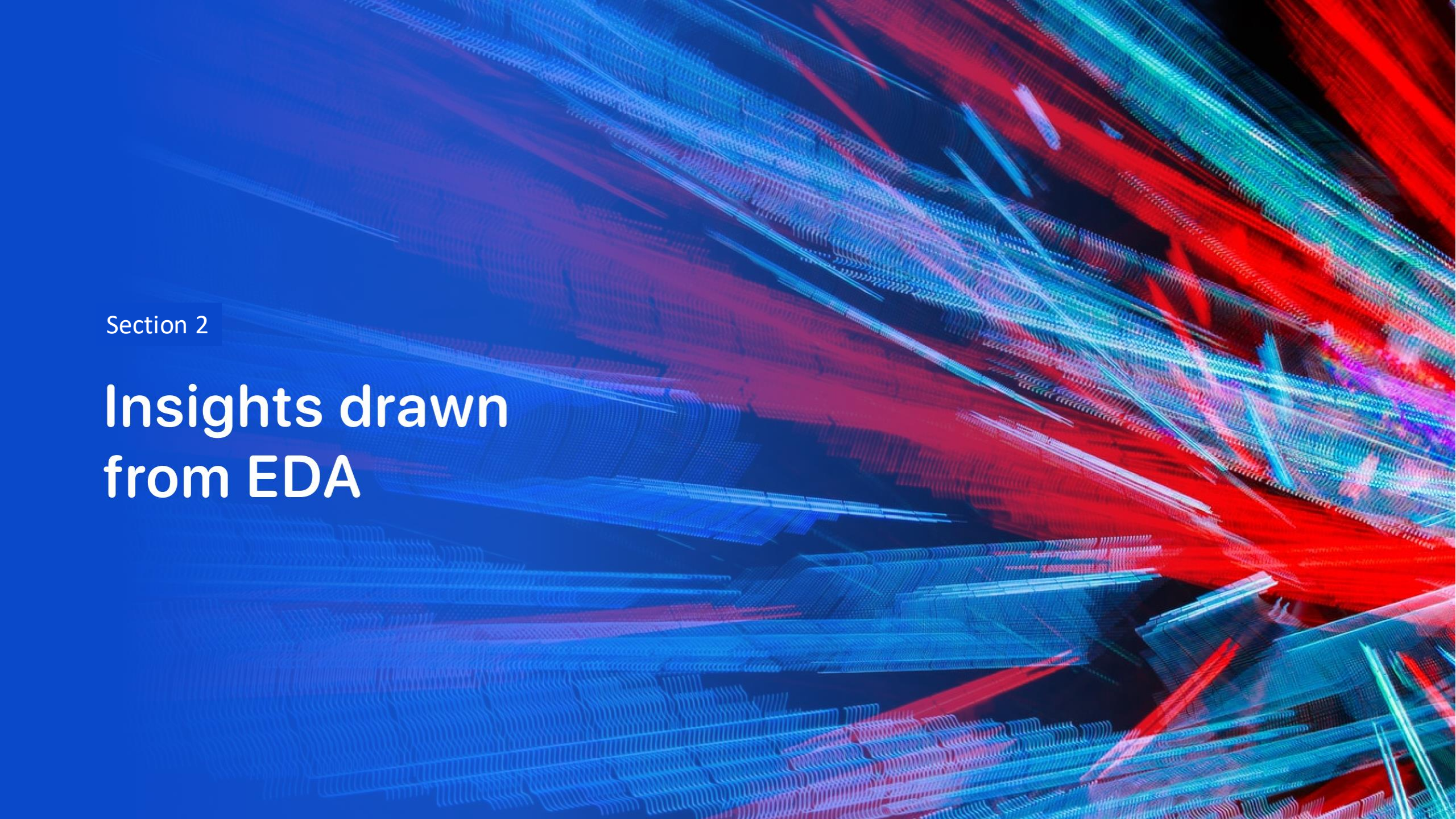
- A drop-down menu that allows you to select the launch site was added. Based on the selection a pie chart of the success rate of the launches from that site was generated.
- A slider for selecting the range of the payload mass was added. Based on the range a scatter chart of the class was generated with colors to indicate the booster version used.
- The GitHub URL of the completed Plotly Dash lab is https://github.com/GHoholis/AppliedDataScienceCapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Initially the data was standardized using .StandardScaler() and the split into training and testing data using the train_test_split.
- A logistic regression object was then created and a grid search was used to find the best parameters for it using the training data.
- The model was then scored on the accuracy obtained on the testing data.
- A similar process was then used for a support vector machine method, a decision tree classifier and a k nearest neighbours method.
- The GitHub URL of the completed predictive analysis lab is https://github.com/GHoholis/AppliedDataScienceCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5_complete.ipynb

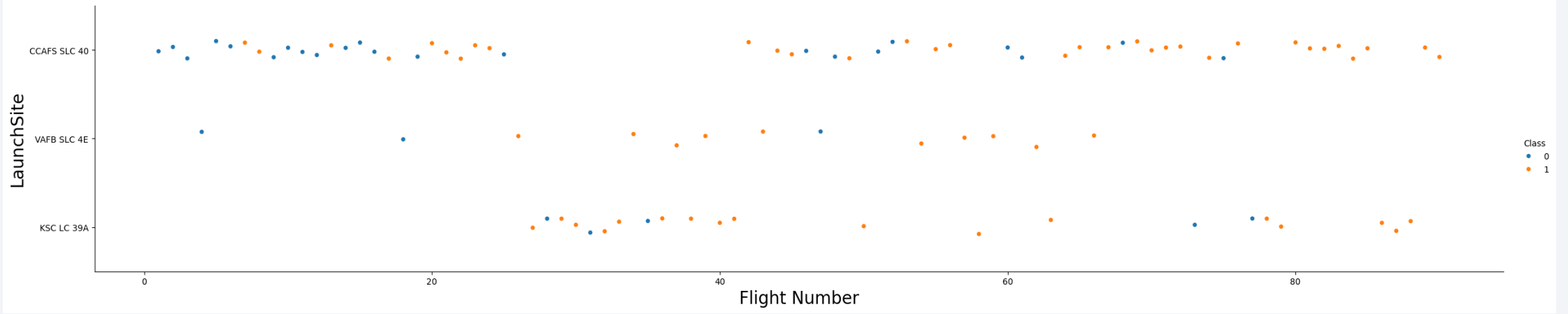
Results

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


Section 2
Insights drawn from EDA

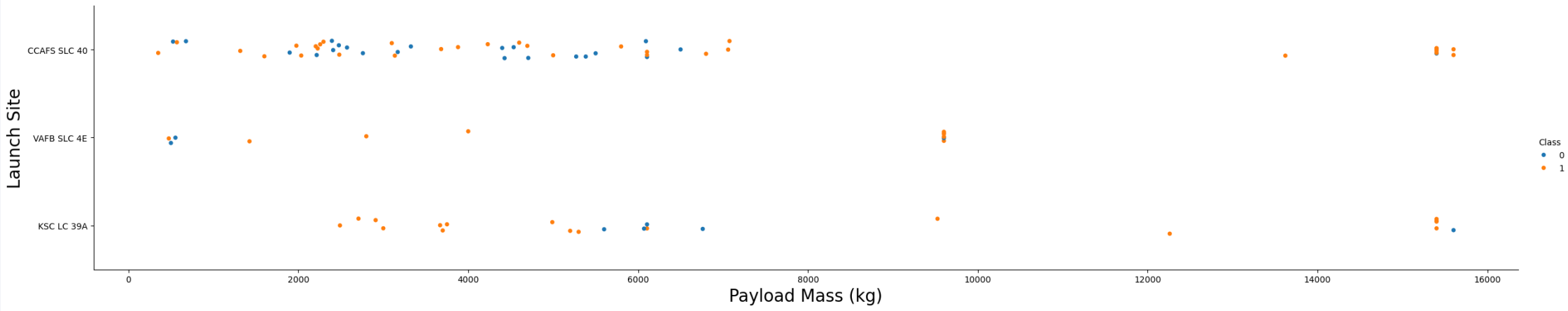
Flight Number vs. Launch Site

- Below is a scatter plot of Flight Number vs. Launch Site.
- The points are colored orange/blue to indicate success/failure, respectively.



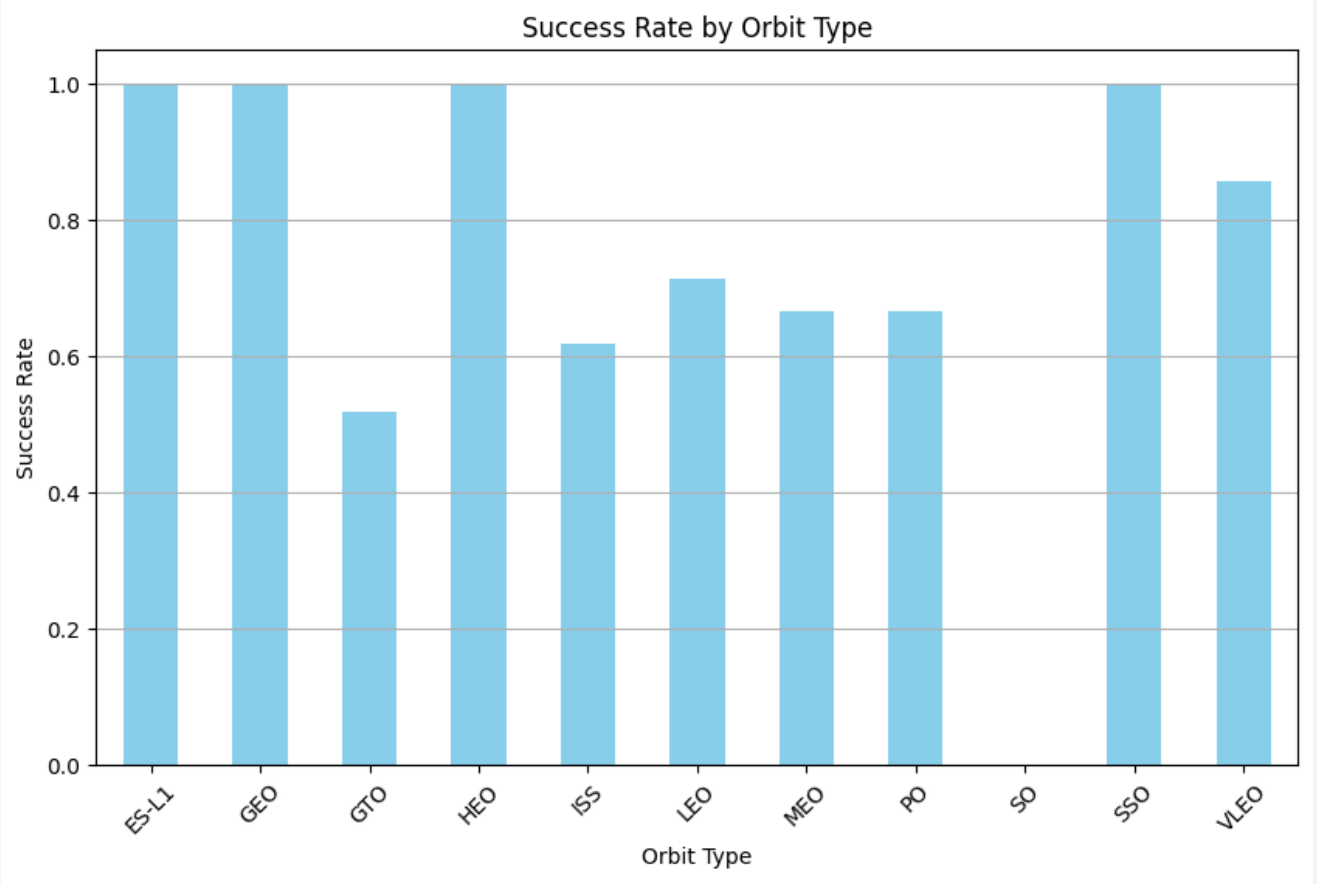
Payload vs. Launch Site

- Below is a scatter plot of Payload vs. Launch Site.
- The points are colored orange/blue to indicate success/failure, respectively.



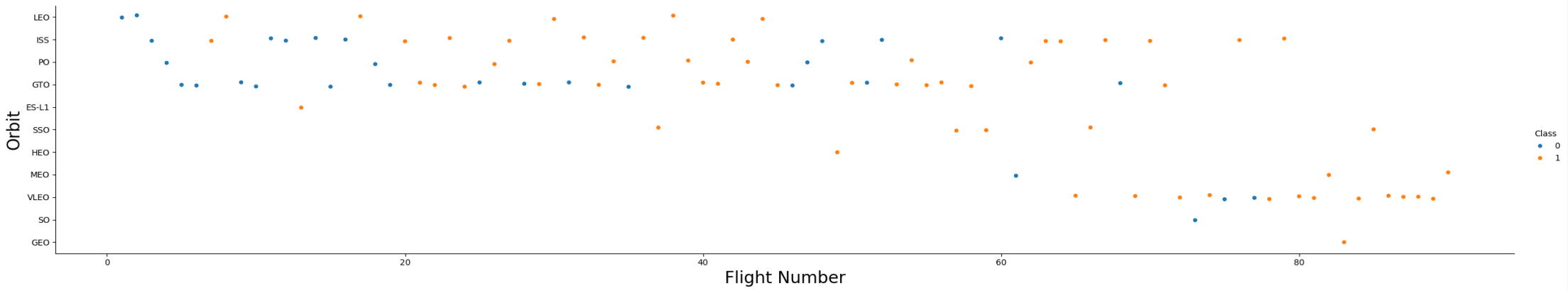
Success Rate vs. Orbit Type

- To the right is a bar chart of success rate vs. orbit type.



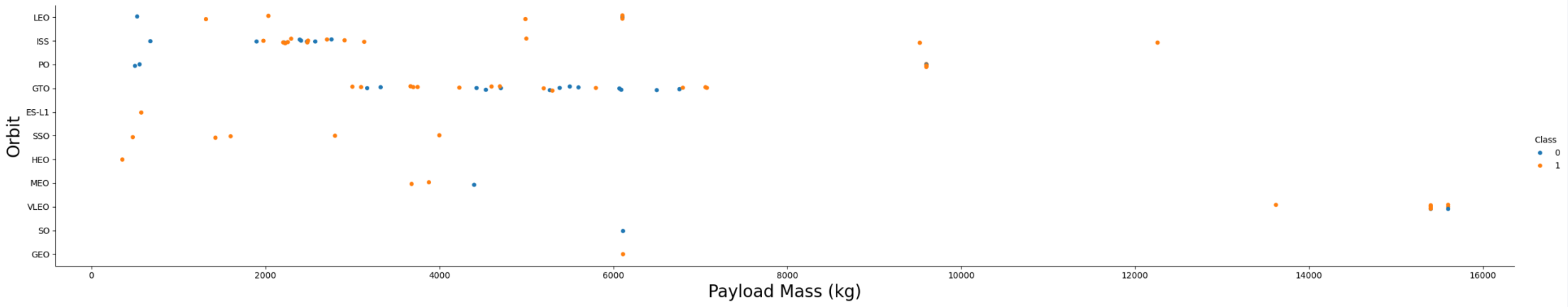
Flight Number vs. Orbit Type

- Below is a scatter point of flight number vs. orbit type
- The points are colored orange/blue to indicate success/failure, respectively.



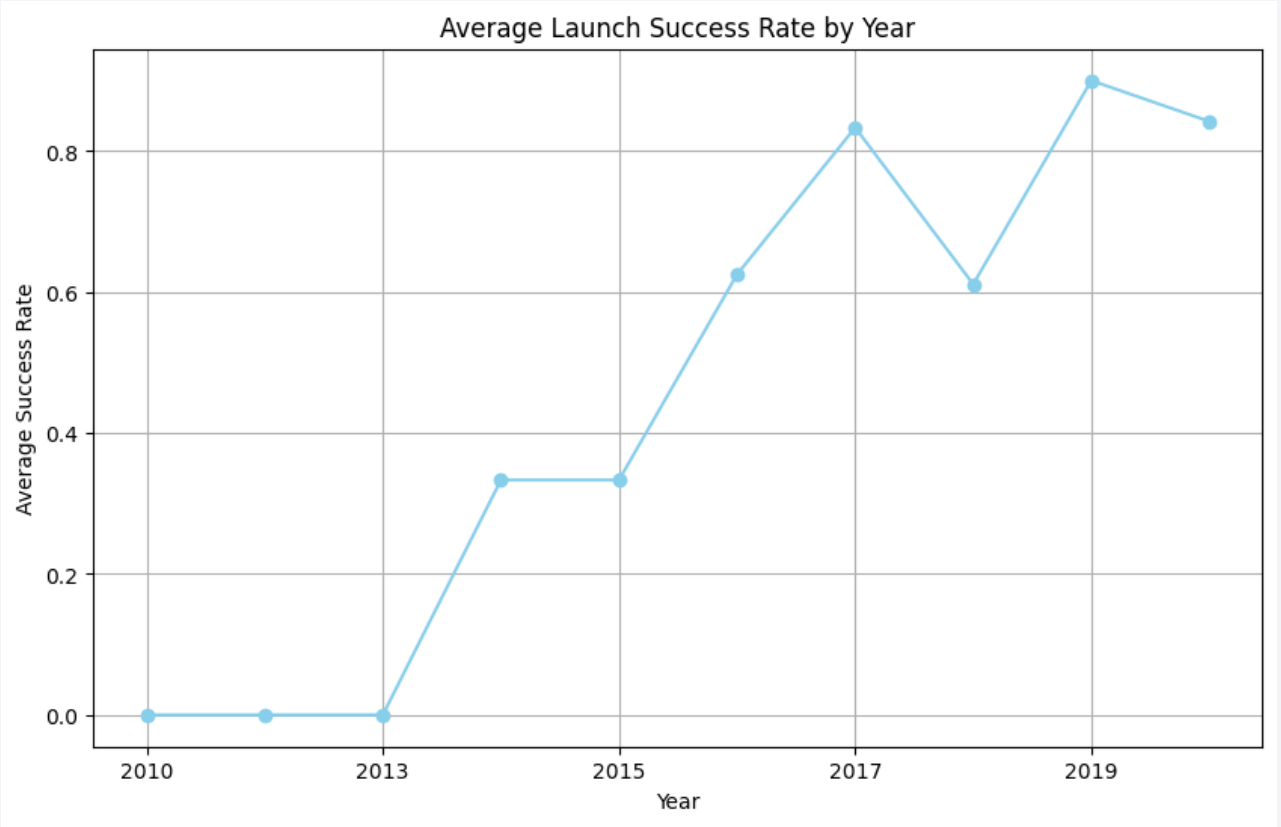
Payload vs. Orbit Type

- Below is a scatter point of payload vs. orbit type.
- The points are colored orange/blue to indicate success/failure, respectively.



Launch Success Yearly Trend

- To the right is a line chart of yearly average success rate.
- The trend observed is that the success rate increases as the years progress.



All Launch Site Names

- The following query was used to obtained the names of the unique launch sites: %sql SELECT DISTINCT "Launch_Site" from SPACEXTBL
- The results can be seen below:

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

Launch Site Names Begin with 'CCA'

- The following query was used to obtain 5 records where launch sites begin with `CCA`: %sql SELECT * from SPACEXTBL where "Launch_Site" like 'CCA%' LIMIT 5
- The results can be seen below:

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

Total Payload Mass

- To calculate the total payload carried by boosters from NASA the following query was used: %sql SELECT SUM("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer"="NASA (CRS)"
- The result can be seen below:

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1 the following query was used: %sql SELECT AVG("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" = "F9 v1.1"
- The result can be seen below:

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- To find the dates of the first successful landing outcome on ground pad the following query was used: %sql SELECT MIN(Date) from SPACEXTBL where "Landing_Outcome" like "Success%"
- The query result can be seen below:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- To obtain a list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 the following query was used: %sql SELECT "Booster_Version" from SPACEXTBL where "Landing_Outcome"="Success (drone ship)" and "PAYLOAD_MASS__KG_">4000 and "PAYLOAD_MASS__KG_"<6000
- The result can be seen below:

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

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failure mission outcomes the following queries were used:
  - %sql SELECT COUNT(*) from SPACEXTBL where "Mission_Outcome" like "Success%"
  - %sql SELECT COUNT(*) from SPACEXTBL where "Mission_Outcome" like "Failure%"
- The results can be seen below:
  - Success:

COUNT(*)
100

  - Failure:

COUNT(*)
1

Boosters Carried Maximum Payload

- To obtain a list of the names of the booster which have carried the maximum payload mass the following query was used: %sql SELECT "Booster_Version" from SPACEXTBL where "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") from SPACEXTBL)
- The results can be seen to the right:

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

- To obtain a list of the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015 the following query was used: %sql SELECT CASE substr(Date, 6, 2) 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, "Booster_Version","Launch_Site","Landing_Outcome" FROM SPACEXTBL WHERE substr(Date, 0, 5) = '2015' AND "Landing_Outcome" LIKE '%failure%' AND "Landing_Outcome" LIKE '%drone ship%';
- The results can be seen below:

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- To obtain the ranks of the counts 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 the following query was used: %sql SELECT "Landing_Outcome", COUNT("Landing_Outcome") AS Outcome_count from SPACEXTBL where Date>="2010-06-04" and Date<="2017-03-20" GROUP BY "Landing_Outcome" ORDER BY Outcome_count
- The results can be seen to the right:

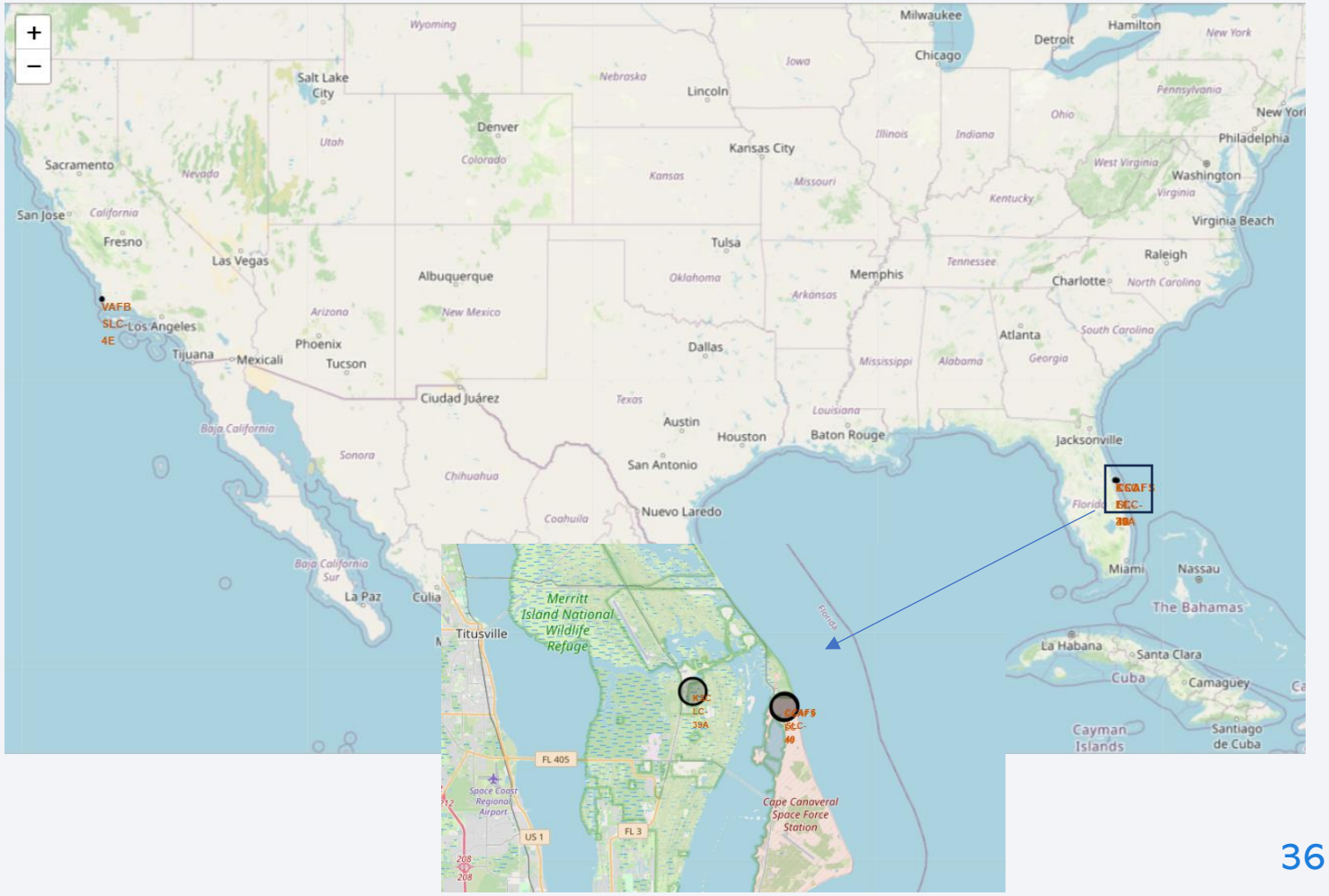
Landing_Outcome	Outcome_count
Precluded (drone ship)	1
Failure (parachute)	2
Uncontrolled (ocean)	2
Controlled (ocean)	3
Success (ground pad)	3
Failure (drone ship)	5
Success (drone ship)	5
No attempt	10

Section 3

Launch Sites Proximities Analysis

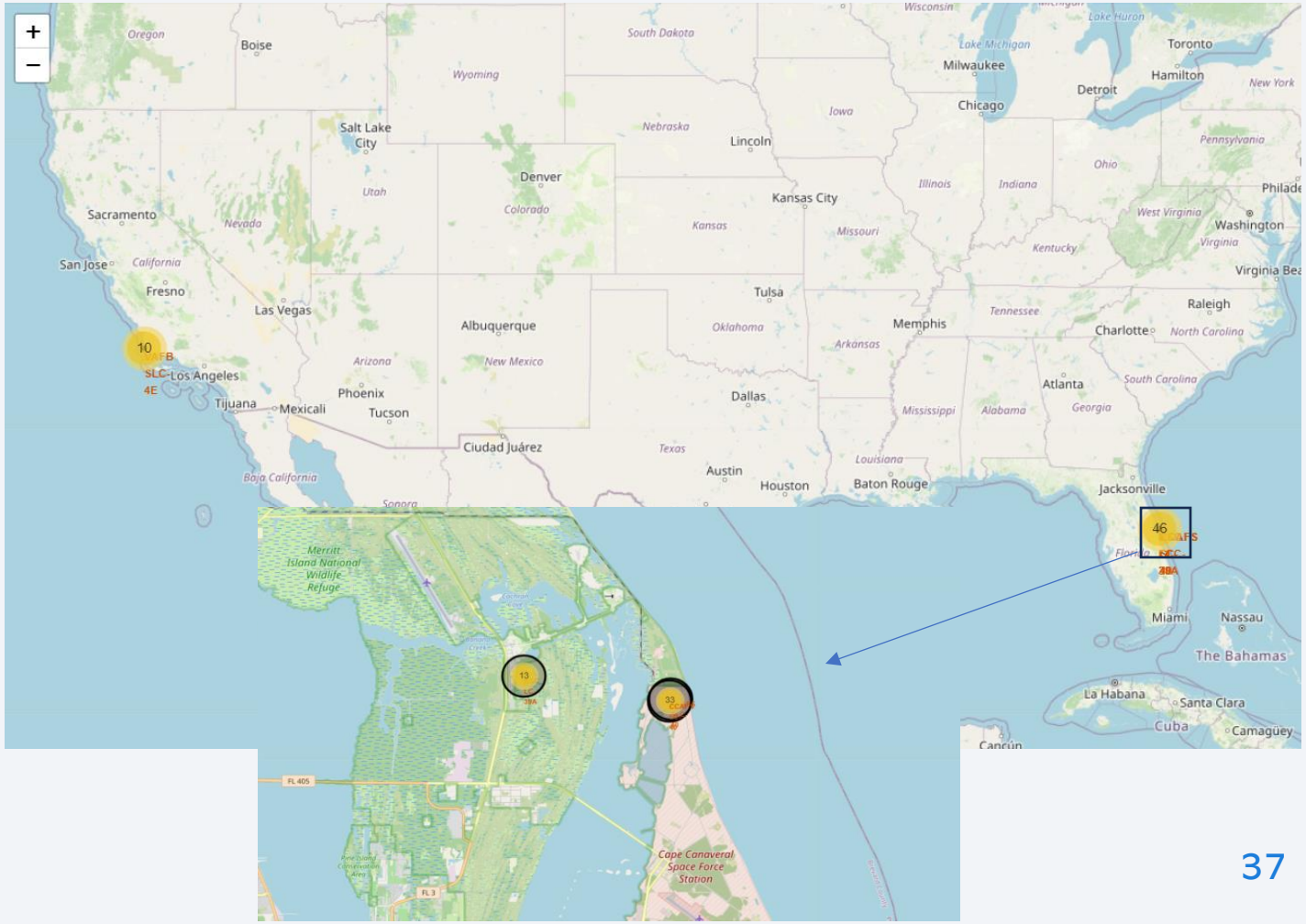
Map Showing Launch Site Locations

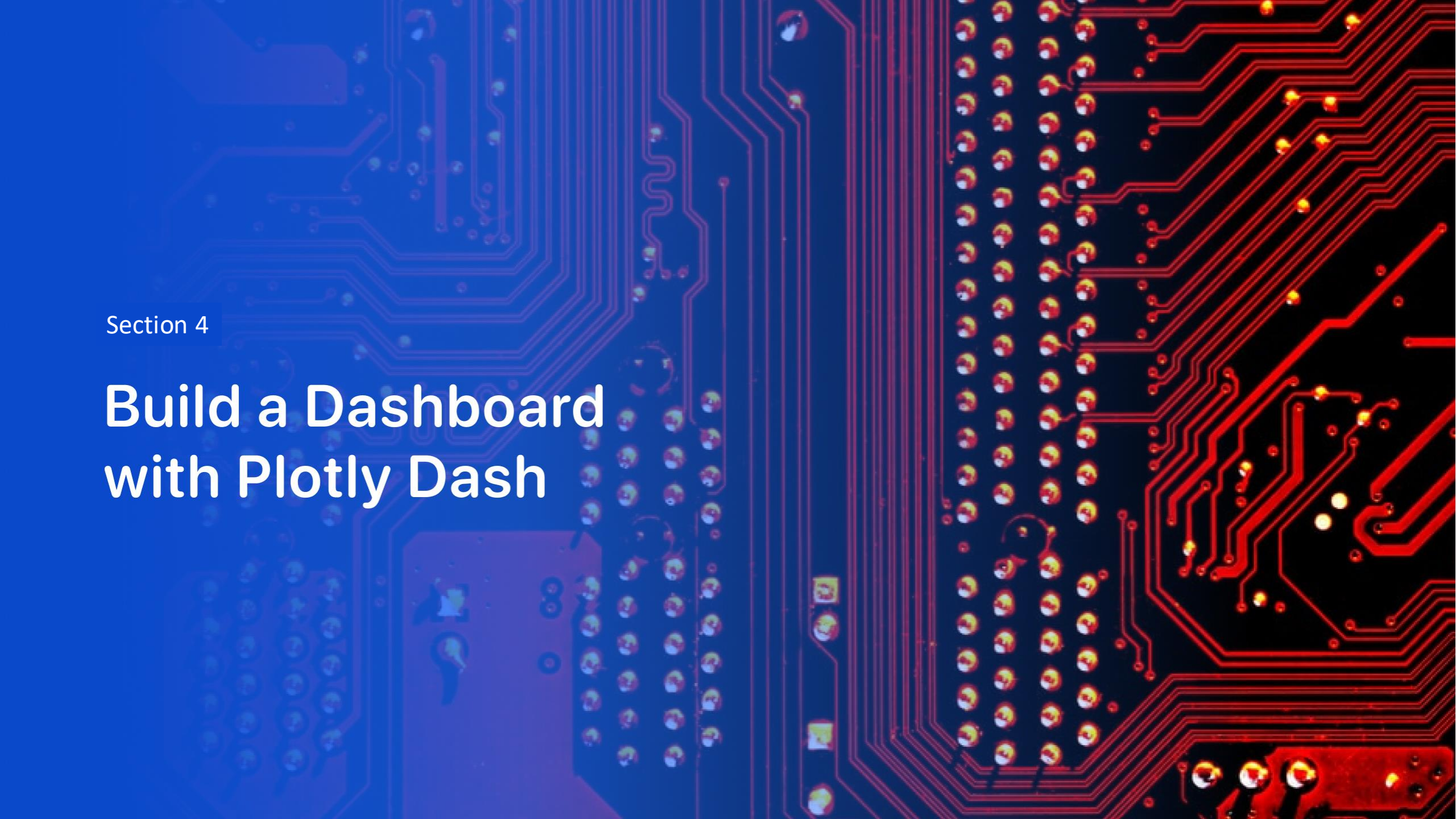
- To the right is a map showing the location of the launch sites.
- As can be seen from the map the launch sites are all located near the coast.



Map Showing Marked with Launches

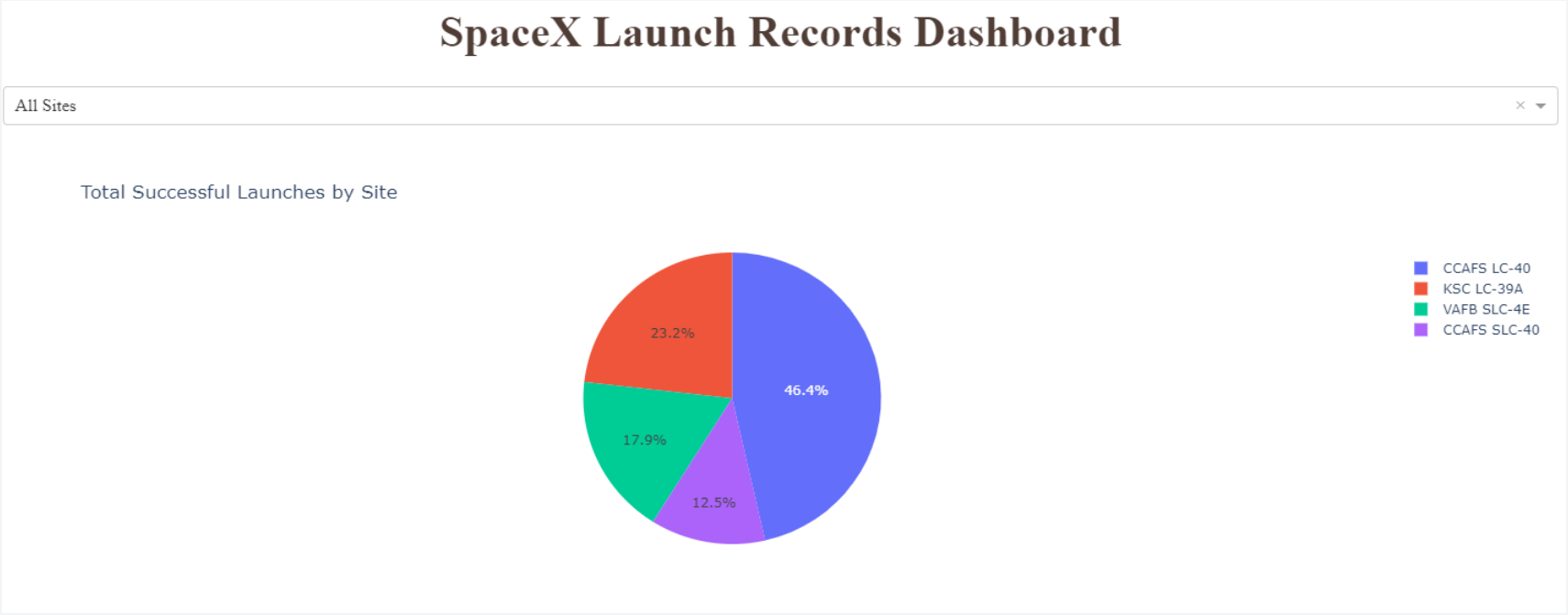
- To right is a map with markers for each of the launches.
- The markers are colored green/red to indicate success/failure, respectively.




Section 4
Build a Dashboard with Plotly Dash

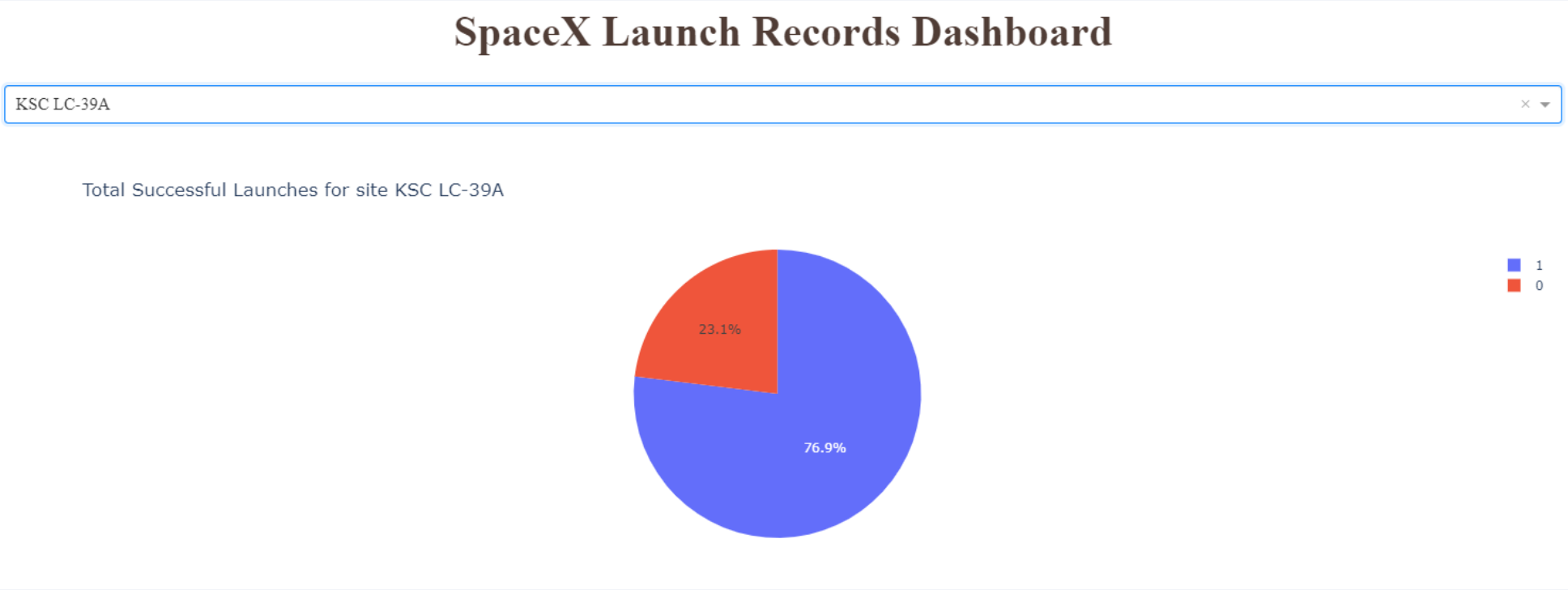
SpaceX Launch Records Dashboard

- Below is a screenshot of launch success count for all sites, using a pie chart.
- As can be seen from the pie chart the most successful launches were from CCAFS LC-40.



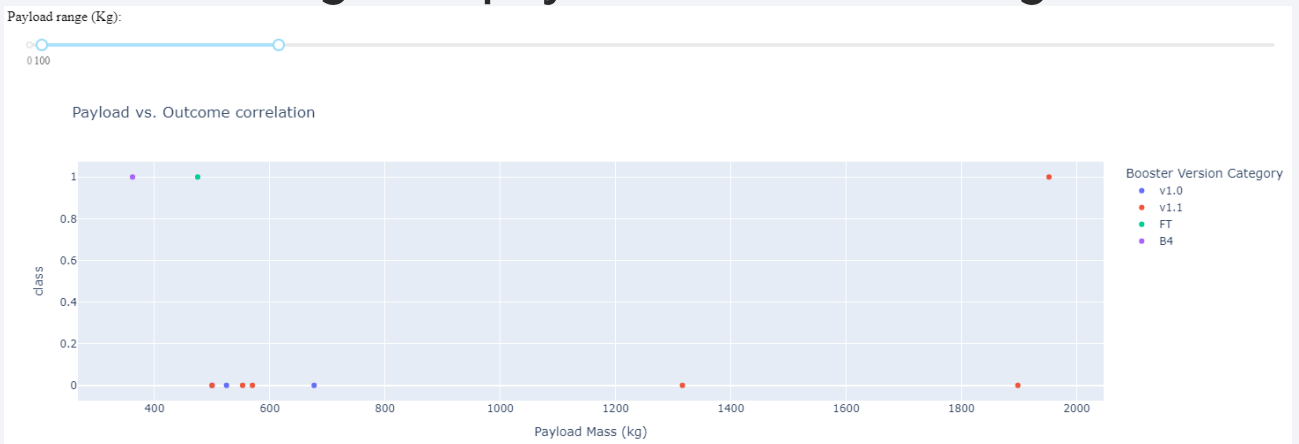
SpaceX Launch Records for KSC LC-39A

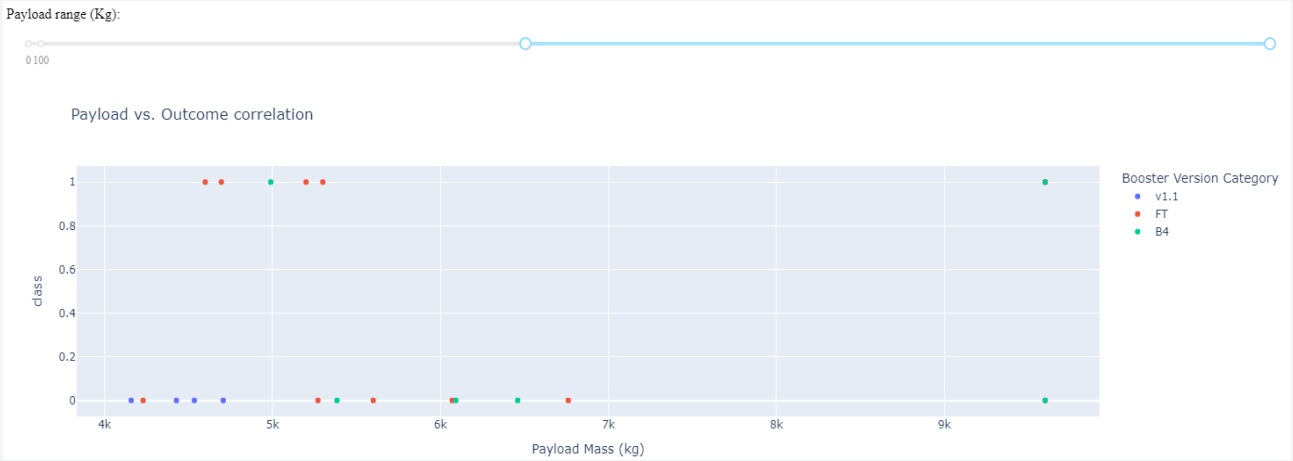
- Below is a screenshot of the pie chart for the launch site with the highest launch success ratio:



Payload vs. Outcome Correlation

- Below are screenshots of Payload vs. Launch Outcome scatter plot for all sites, with payloads under 2000kg and payloads over 4000kg:



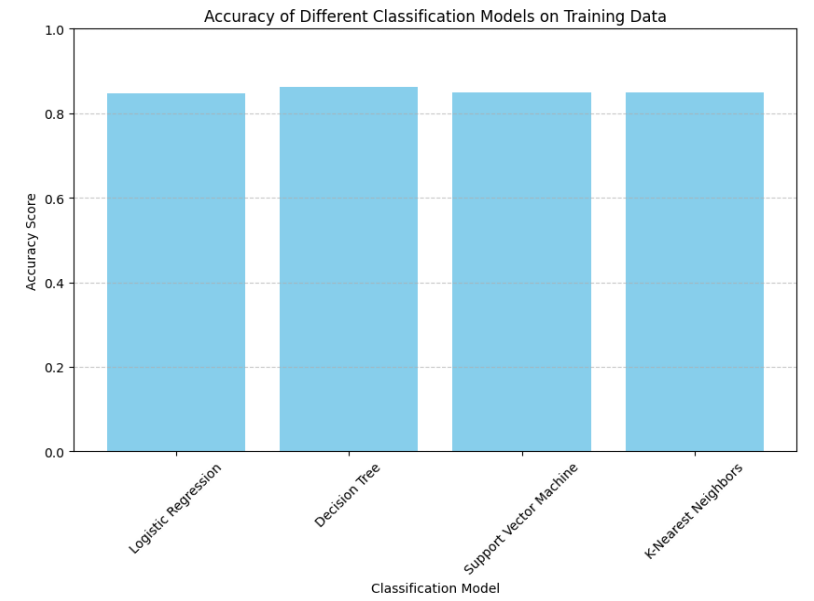


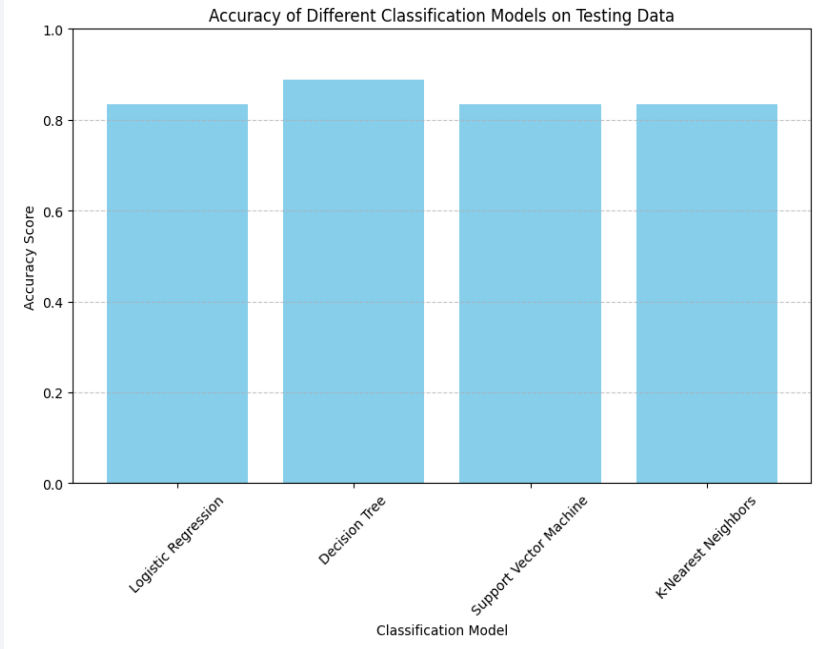
Section 5

Predictive Analysis (Classification)

Classification Accuracy

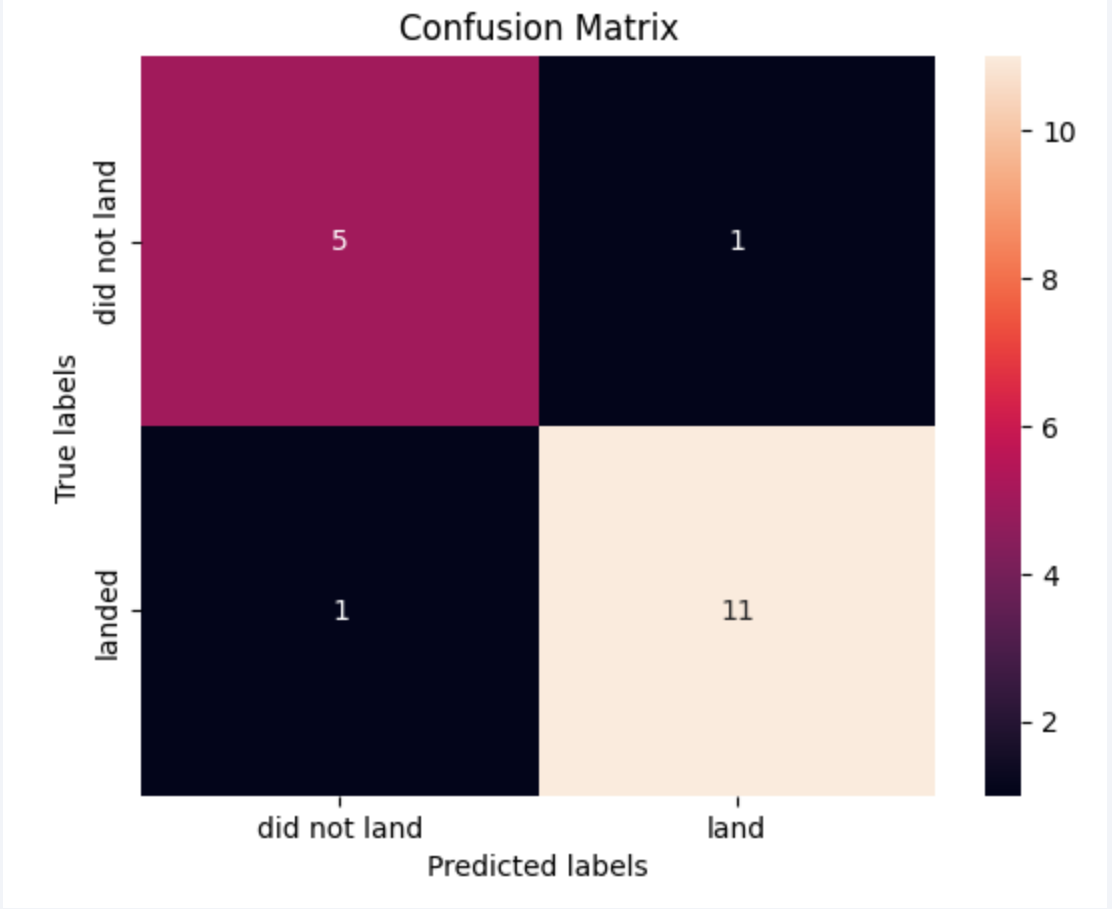
- Bar charts showing the built model accuracy for all built classification models can be seen below.
- The decision tree classifier has the highest classification accuracy.





Confusion Matrix

- The confusion matrix of the decision tree classifier can be seen to the right.
- As can be seen from the plot there is one false positive and one false negative.
- There are five true negatives and eleven true positives making this the best model in terms of prediction accuracy.



Conclusions

- By mapping the SpaceX launch sites and creating interactive visualizations, we gained insights into the geographical distribution and performance of different launch sites. The analysis of launch success rates across various sites showed patterns that could inform future launch planning and site development.
- The creation of a Dash application with dropdowns, pie charts, and scatter plots provided a user-friendly interface for exploring SpaceX launch data. Users can now easily visualize the success and failure rates of launches, both in general and for specific sites, facilitating better decision-making based on historical data.
- Building various classification models (e.g., Logistic Regression, Decision Tree, SVM, K-Nearest Neighbors) allowed us to predict the success of SpaceX launches with varying degrees of accuracy. The model accuracy comparison revealed that the Decision Tree model performed the best.

Thank you!