IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zach Wittl
9/29/23

Outline

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

Executive Summary

- Summary of Methodologies
  - Data Collection and Wrangling
  - Exploratory Data Analysis and Interactive Visual Analytics with SQL
  - Interactive Map with Folium and Plotly Dash
  - Predictive Methodology and Analysis
- Summary of all results

Introduction

- Project Context
  - I am a Data Scientist working for SpaceY, a SpaceX competitor that is owned by Allon Mask. This project will provide data driven insights that determine whether SpaceX will reuse the first stage of a Falcon 9 rocket.
  - If we can successfully determine whether the first stage will land then we can determine the cost of a SpaceX launch. This information can be used by SpaceY to make a competitive bid against SpaceX.
- Problems to Solve
  - Determine the price of each launch.
  - Will Space X reuse the first stage of a Falcon 9 rocket? Will it land successfully to be reused?
  - Use machine learning model to predict whether Space X will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

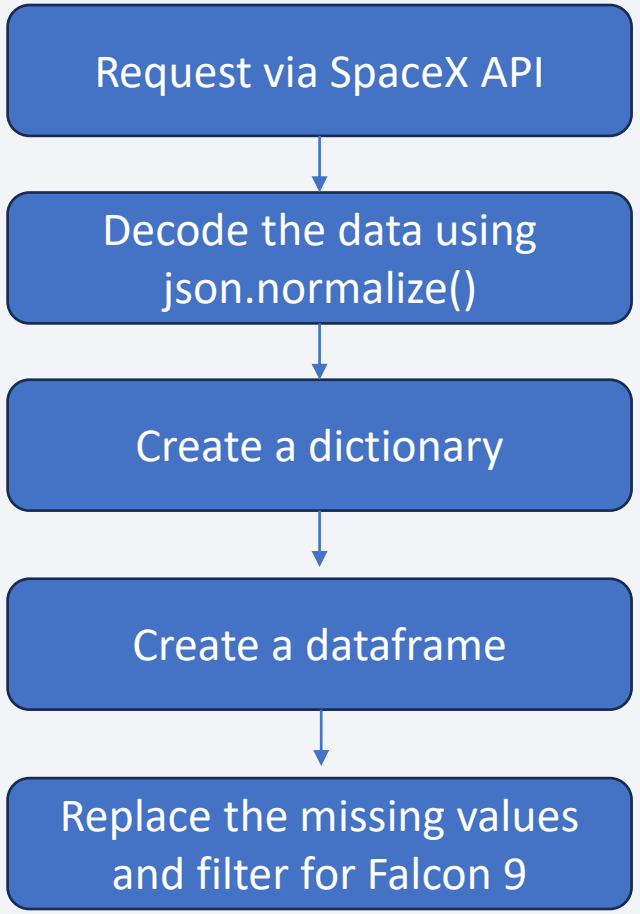
- Data collection methodology:
  - Collected Space X data via Space X REST API and Launch Data via Web Scraping with BeautifulSoup.
- Perform data wrangling
  - Convert the JSON file from Space X REST API into a data frame.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
  - The Data was collected via two means.
    1. Space X Rest API to gather the Rocket Launch information
    2. Data Scraping via BeautifulSoup to collect the Falcon 9 launch data via Wikipedia

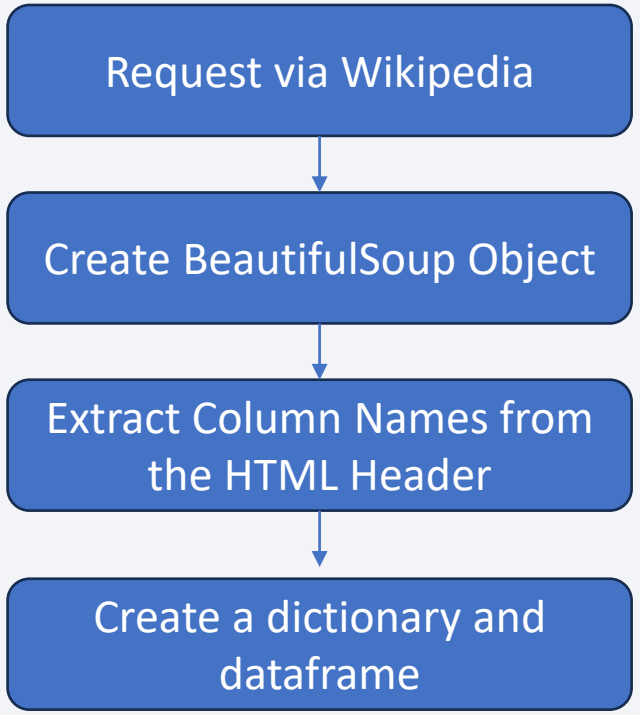
Data Collection – SpaceX API

- In Jupyter notebook, collect data regarding the SpaceX launches by requesting from the SpaceX API.
- Normalize the data and covert it into a dataframe.
- Filter the data to only include launch information for the Falcon 9 rocket and replace any missing values with the mean.
- Github Link



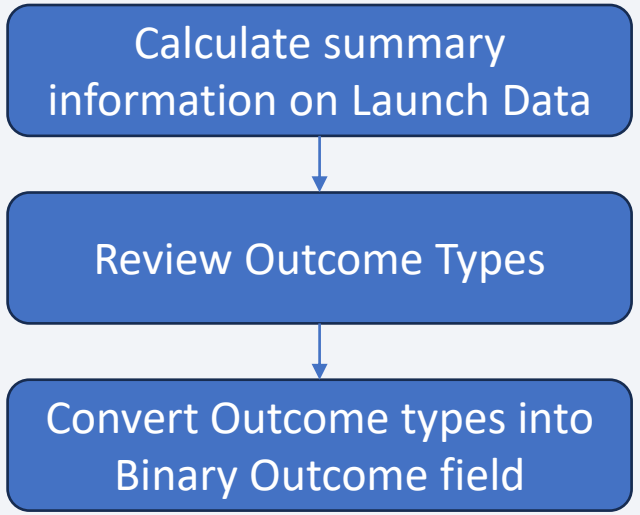
Data Collection – Scraping

- In Jupyter notebook, collect data regarding the SpaceX launches by pulling from Wikipedia.
- Create a BeautifulSoup Object and Extract Column information from the HTML Header.
- Create a dictionary and a dataframe to perform analysis.
- Github Link



Data Wrangling

- Data Wrangling Process
  - Prepare for Exploratory Data Analysis
  - Calculate:
    - The number of launches at each launch site
    - The number of launches for each orbit
    - The number of mission outcomes for each orbit
  - Create a landing outcome label based on the outcome column



- [Data Wrangling Github Link]

EDA with Data Visualization

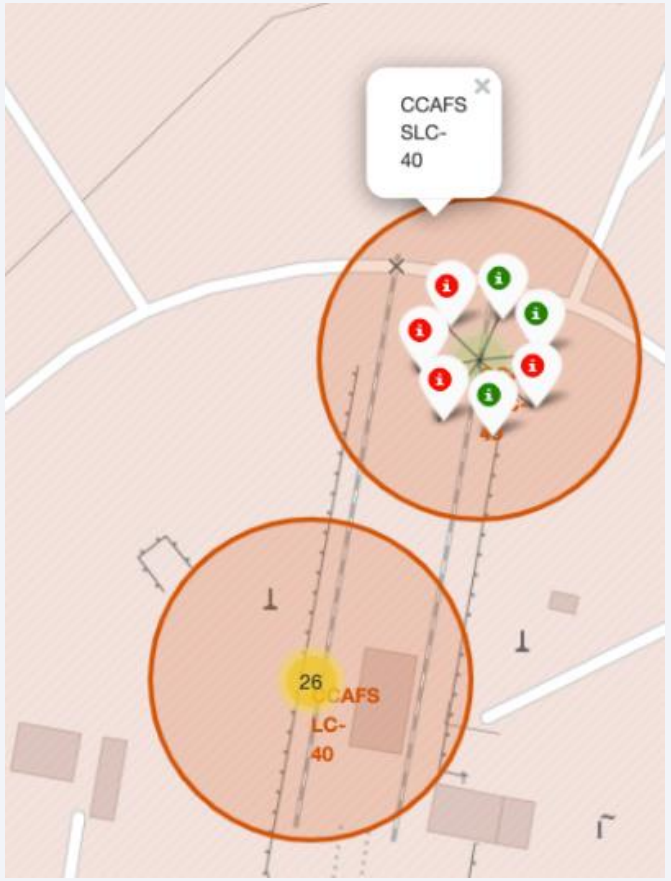
- After collecting and wrangling the SpaceX data the next step is visualize the data to make it easier to analyze and identify patterns.
- Charts Created:
  - Flight Number vs Payload Amount
  - Flight Number vs Launch Site
  - Payload Mass va Launch Site
  - Payload Mass vs Orbit Type
- It was important to view the variables in scatter plot to see if there was a linear relationship.
- Data Visulization Github Link

EDA with SQL

- SQL queries Executed to Explore the Data:
  - Select distinct launch sites.
  - Pull data where the launch site begins with ‘CCA’.
  - Total Payload Mass when the boosters used were launched by NASA.
  - Average Payload Mass for the Falcon 9 Booster.
  - Date of first successful Ground Pad landing.
  - Total number of missions by outcome type.
- SQL Data Analysis Github Link

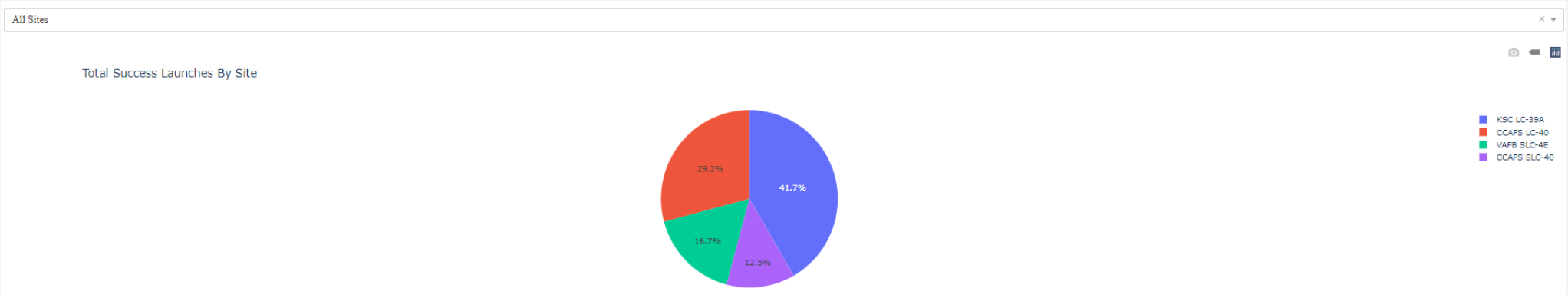
Build an Interactive Map with Folium

- Marked Launch Sites
  - Added Blue Circle to NASA Johnson Space Center.
  - Added Red Circle to all Launch Sites with popup information containing launch site coordinates.
- Launch Outcomes
  - Added color markers to highlight the outcome of a launch. Green for successful and red for launch failure.
- Distance between Launch Sites
  - Added marker lines to display distance between launch sites and how close launch sites were to coastline.
- Visual Analytics with Folium Github Link



Build a Dashboard with Plotly Dash

- Created Visual Dashboard that contained drop list that allowed user to filter by Launch Site.
- Created Pie Chart that displayed the number of Successful launches by Launch Site.
- Installed slider that allowed user to select launches based on Payload Mass.
- Created Scatter chart that displayed impact of Payload Mass vs. Success rate by Booster Version.
- Visual Analytics with Plotly Github Link

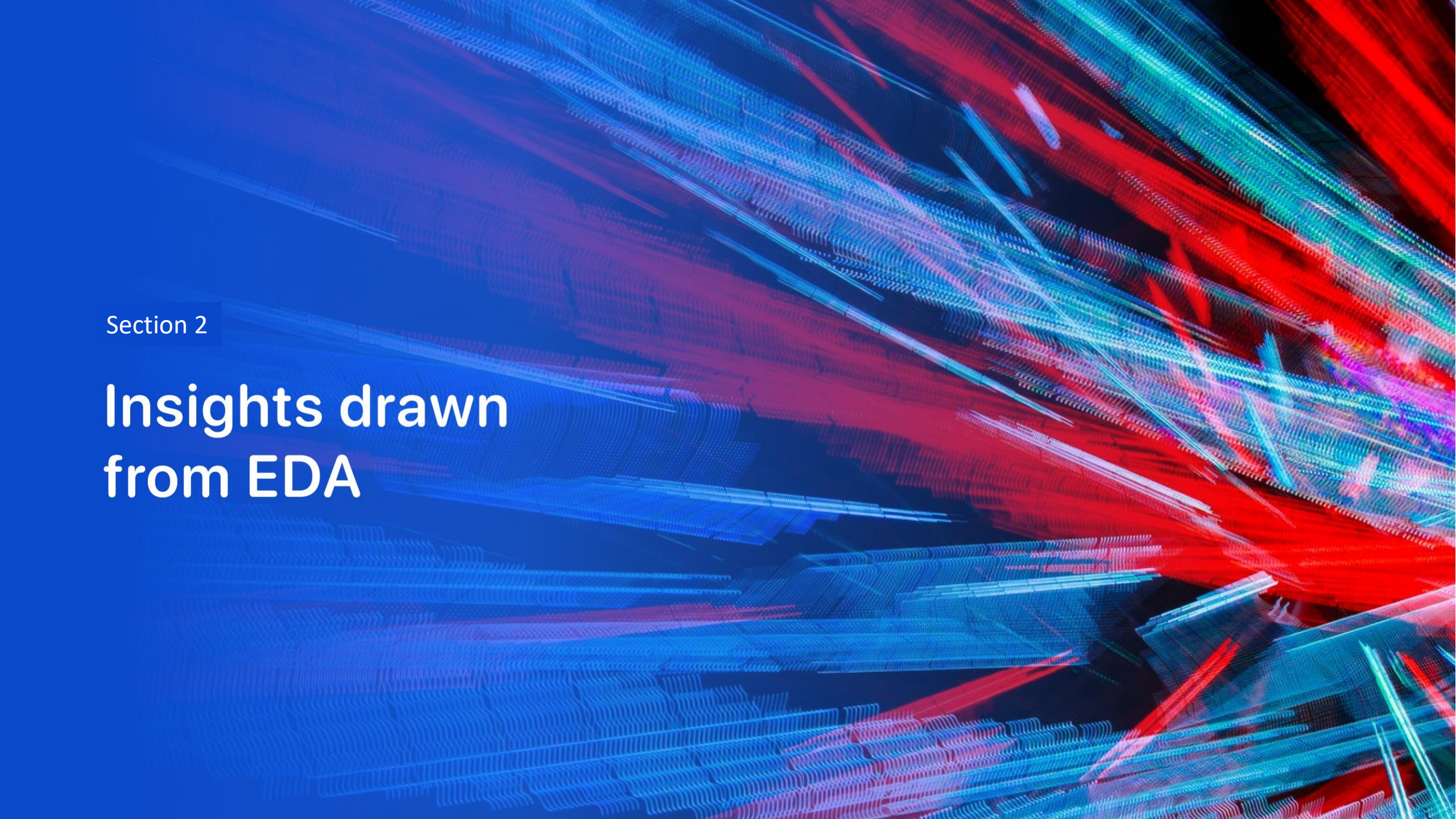


Predictive Analysis (Classification)

- First step was to create a **chart** using NumPy.
- Then we needed to **standardize** the data with StandardScaler.
- After data was prepped we then **split** the data into training and test data sets.
- We **optimized** the parameters by using the GridSearchCV and set it = 10.
- Applied the GridSearchCV on different modeling approaches such as Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbor.
- Determined the **accuracy** of each model to calculate most effective approach for prediction.
- Analyzed the **confusion matrix** for each model to help with decisioning.
- **Selected the best model** based on Jaccard Score, F1 Score, and Accuracy measurements.
- Machine Learning Predictive Analysis Github Link

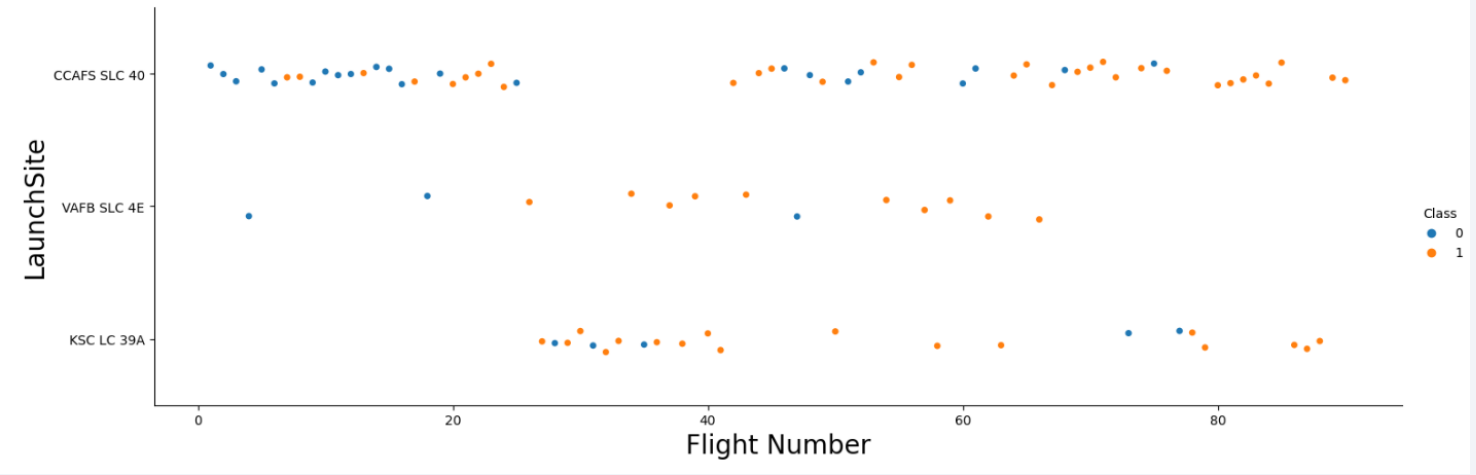
Results

- Exploratory data analysis results
  - Falcon 9 Launch Success has improved over time.
  - The most successful launch site is KSC LC-39A.
- Interactive analytics demo in screenshots
  - Launch sites are positioned near the coast and in the southern part of the country.
- Predictive analysis results
  - The best model for predictive analysis is the decision tree model based on Jaccard Score, F1 Score, and Accuracy measurements.


Section 2
Insights drawn from EDA

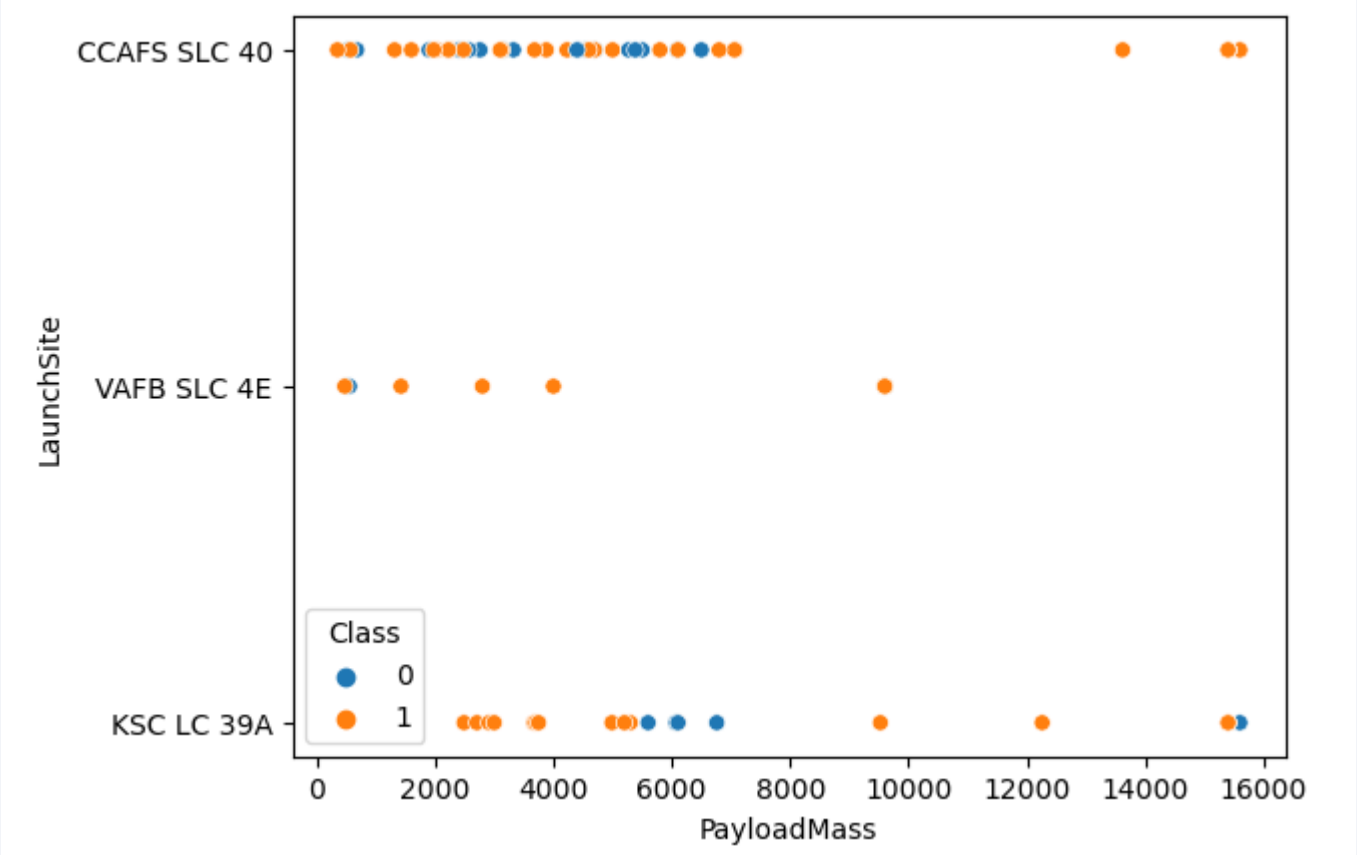
Flight Number vs. Launch Site

- The scatter plot displayed Flight Number by Launch Site.
- The Colors indicate whether the Launch Failed or Succeeded. Blue being Fail and Orange Success.
- More failures (Blue) occurred at earlier launches.



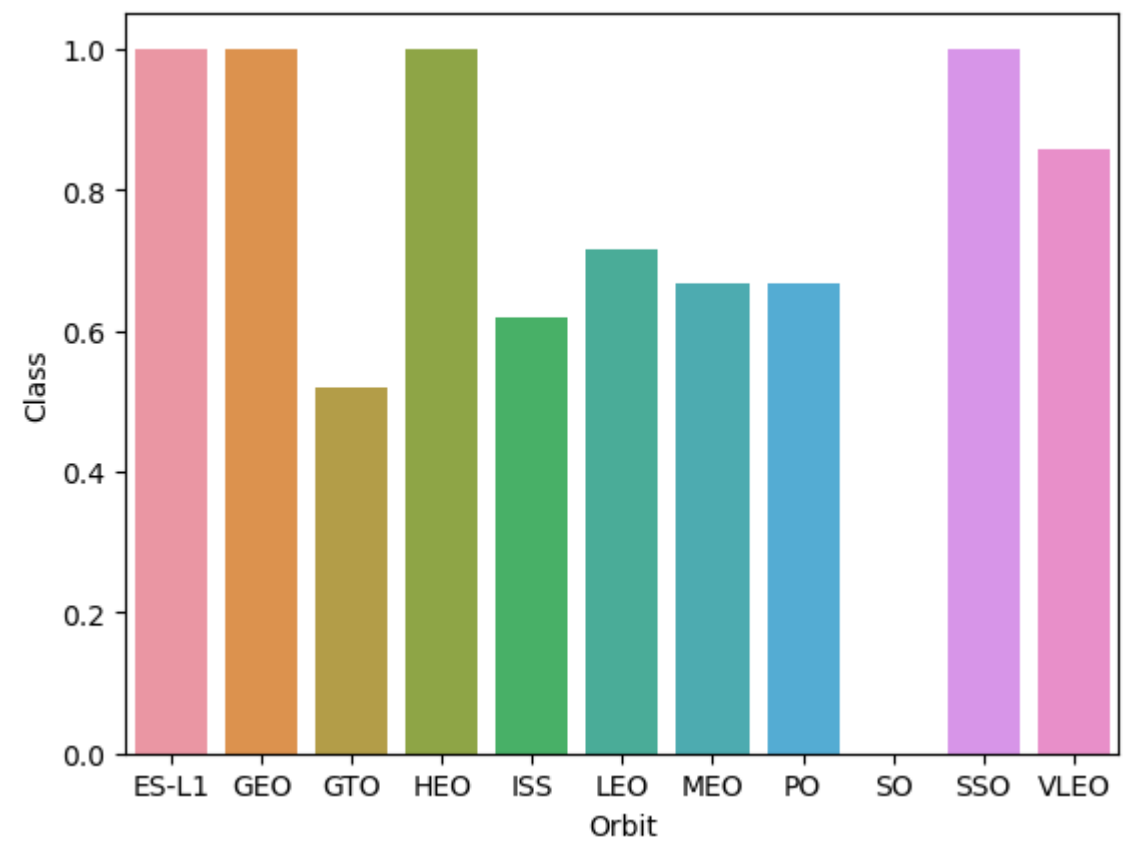
Payload vs. Launch Site

- The higher the payload Mass the more successful the launch was.
- Most launches with a Payload Mass > 7k were successful.



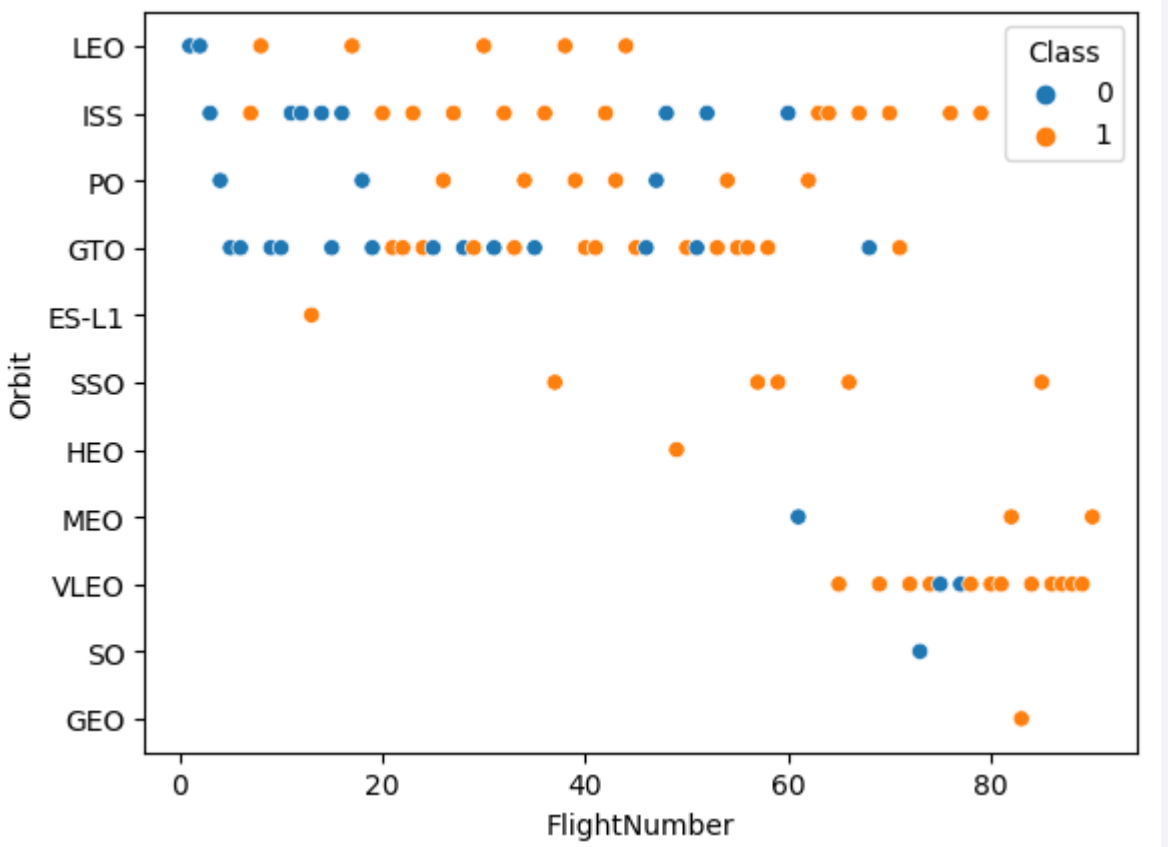
Success Rate vs. Orbit Type

- ES-L1, GEO, and HEO had a 100% success rate for Launches.
- SO was the worst performing launch type with a 0% success rate.



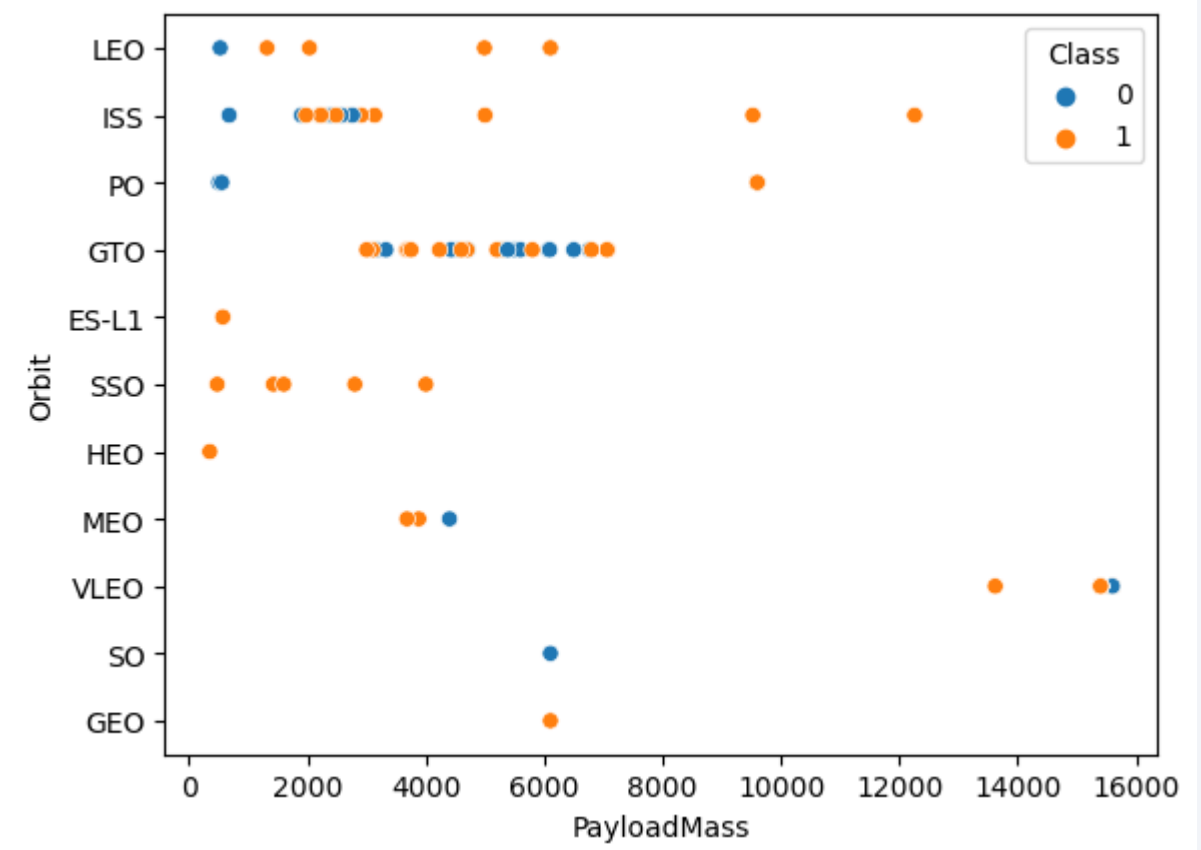
Flight Number vs. Orbit Type

- Generally, the more flights that occur for an Orbit type, the more successful the launches are.



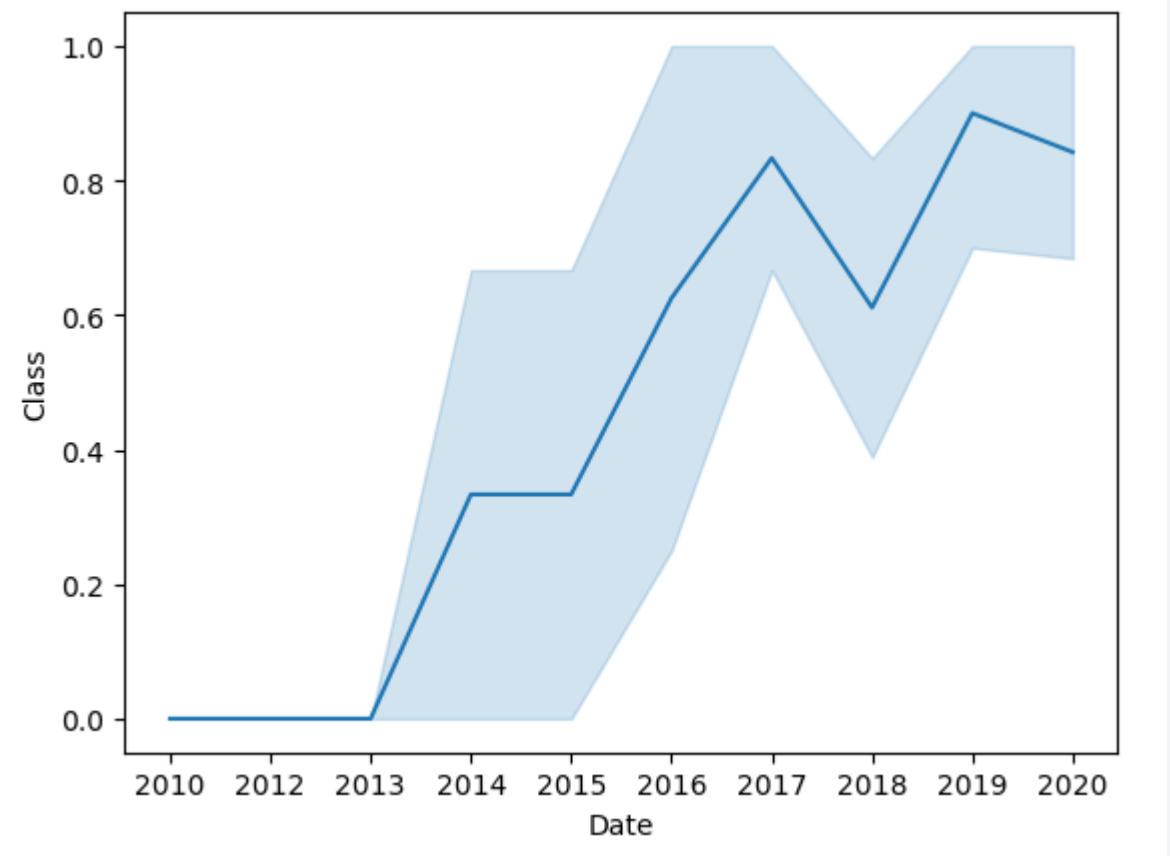
Payload vs. Orbit Type

- Heavier Payloads are successful with the ISS and PO Orbit Type.
- GTO orbit type had mixed results regardless of Payload amount.



Launch Success Yearly Trend

- Initial results had a low success rate.
- Success rate increased over time, peaking in 2018.



All Launch Site Names

- This query pulled the 5 distinct launch Sites.

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

Launch Site Names Begin with 'CCA'

- This query was used to pull all Launch Sites that begin with CCA

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

```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- This query summed up the Payload Mass field to generate the Total Payload Mass amount.

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

```
 * sqlite:///my_data1.db
Done.
```

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

- This query calculated the average payload mass amount where the Booster Version was equal to F9 v1.1.

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

```
 * sqlite:///my_data1.db
Done.
```

AVERAGE_PAYLOAD_MASS
2928.4

First Successful Ground Landing Date

- The first successful Ground Landing Date occurred on December 12th, 2015.

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL \
    WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

FIRST_SUCCESSFUL_GROUND_LANDING
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query provided the 4 Booster Versions that were able to successfully land on a Drone Ship with a payload between 4k and 6k kgs.

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
    WHERE (LANDING_OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- This query calculates the total number of Mission Outcomes.

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

- This query lists the Booster Versions that have carried the Maximum Payload Amount from the table.

```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL \
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

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

- This query lists the failed landing_outcomes in drone ship, their booster versions, and launch site names for the year 2015.

```
%sql SELECT substr(date,6,2) as Month, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME, date FROM SPACEXTBL \
    WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND substr(date,1,4) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

- This query ranks 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(LANDING_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY LANDING_OUTCOME \
    ORDER BY TOTAL_NUMBER DESC;
```

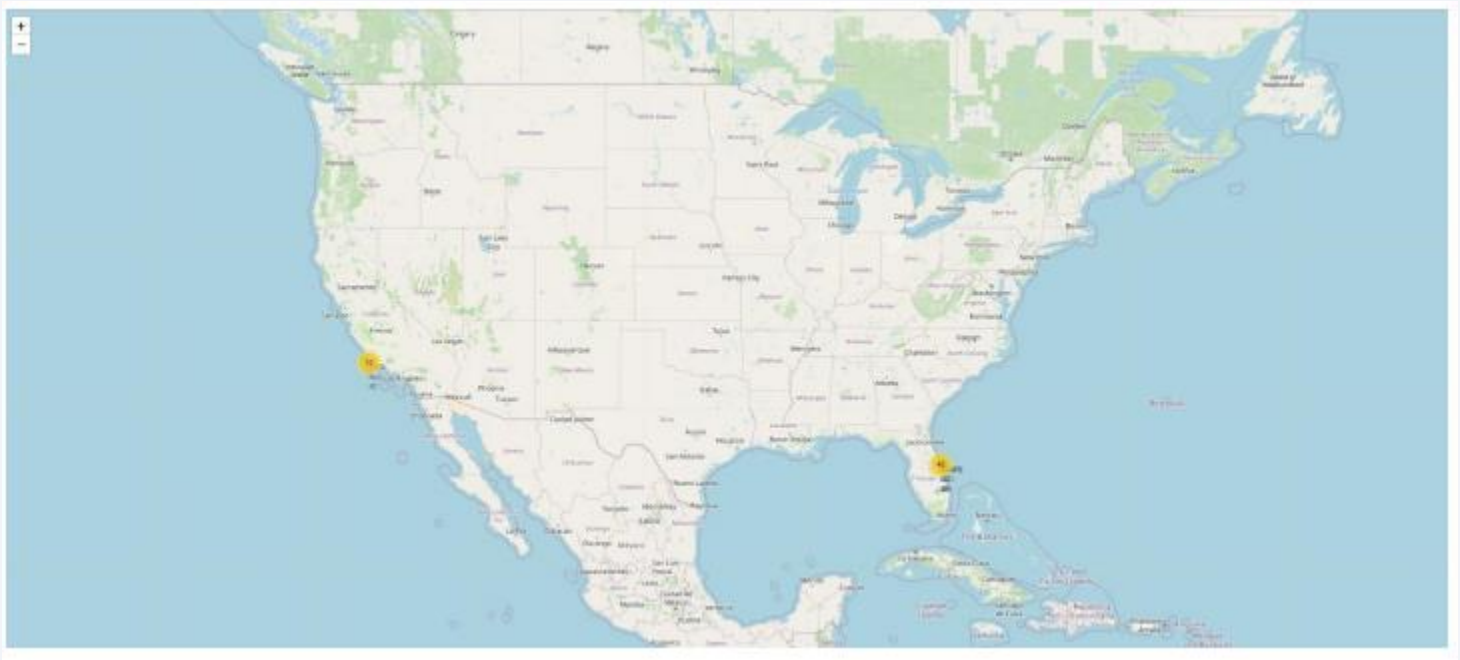
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

Launch Sites Proximities Analysis

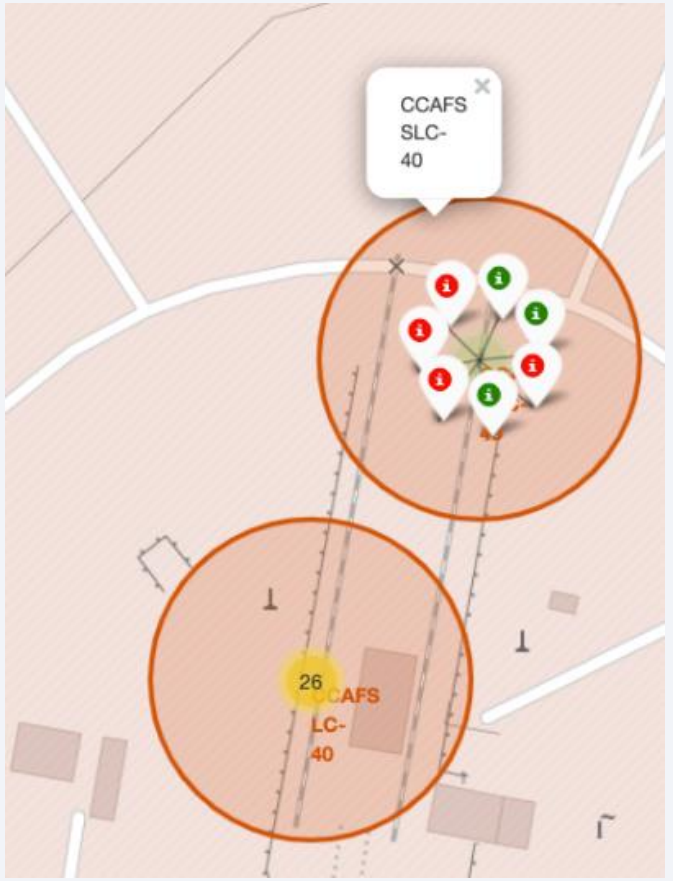
Folium Global Launch Site Markers



- Launch Sites are placed near coastlines in the southern regions of the country.

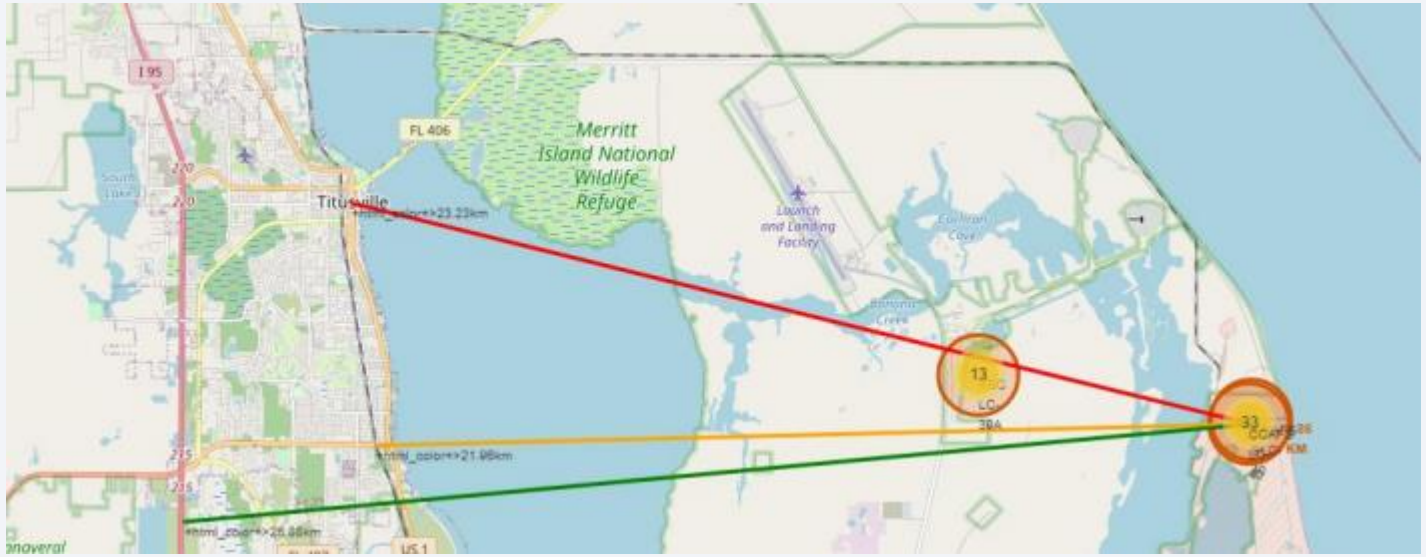
Folium Launch Site Outcome Visualization

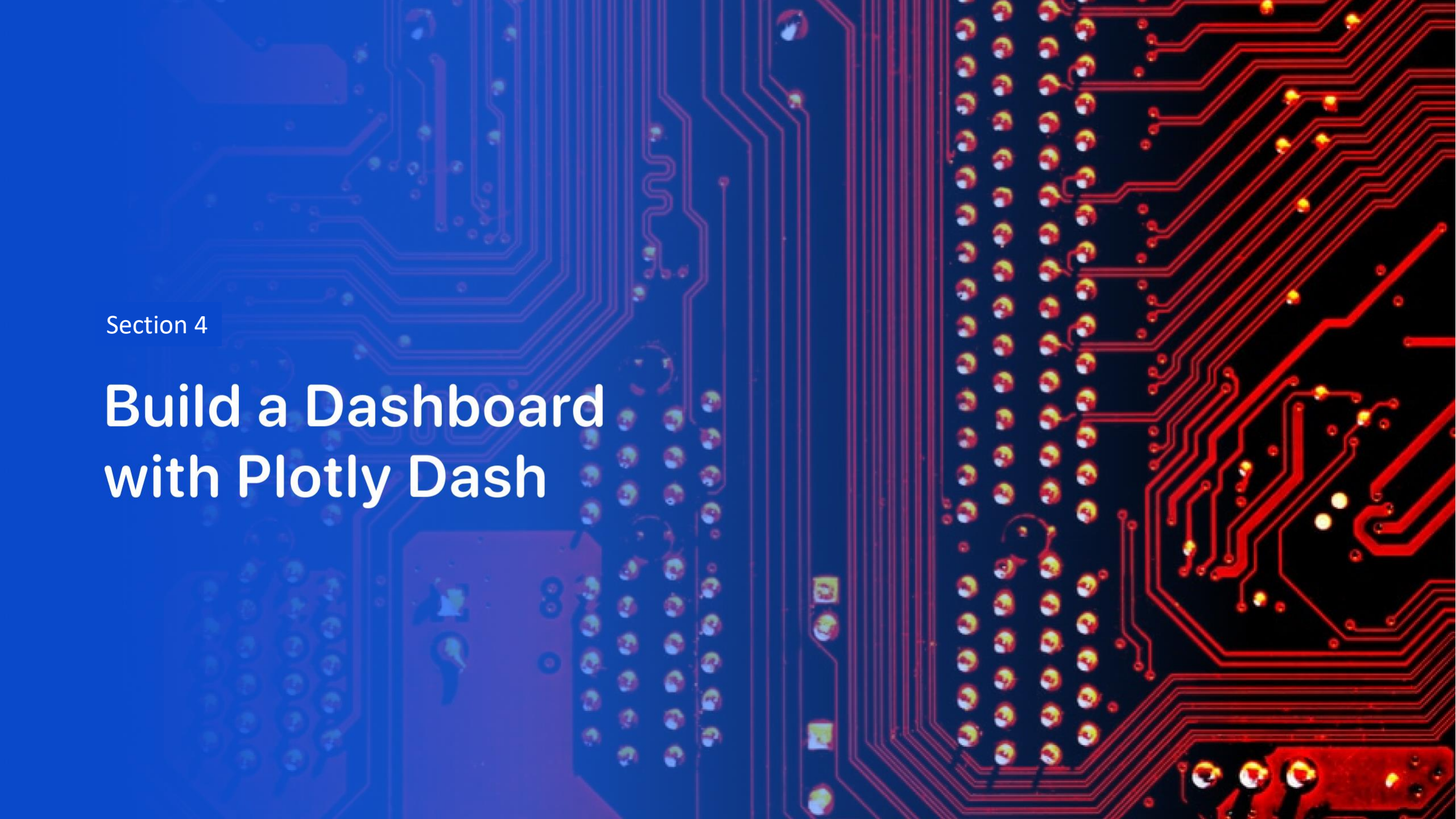
- Green is used to indicate a successful launch.
- Red indicates the launch was a failure.
- Launch Site CCAFS SLC-40 has had 3 successful launches and 4 failed launches.



Folium Proximity to nearby Infrastructure

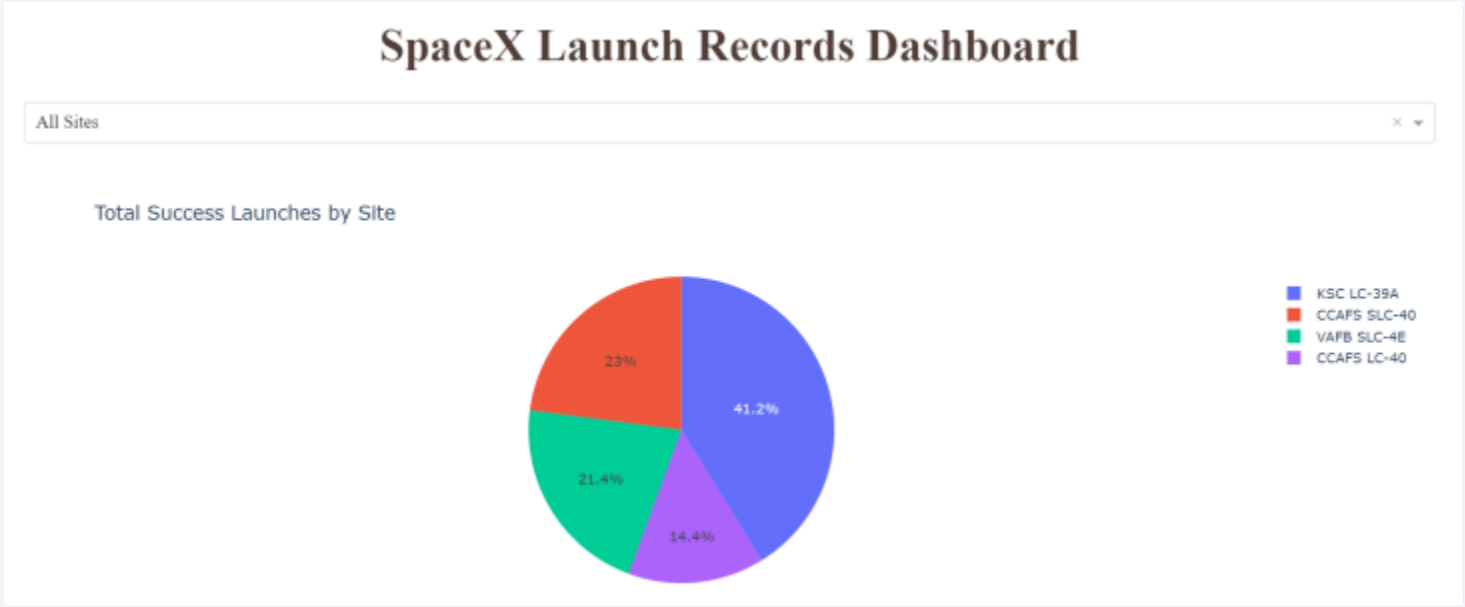
- The proximity lines indicate how far away the launch sites are from critical infrastructure. These launch sites are on the coast line and away from highways and cities to minimize impact of a failed launch.




Section 4
Build a Dashboard with Plotly Dash

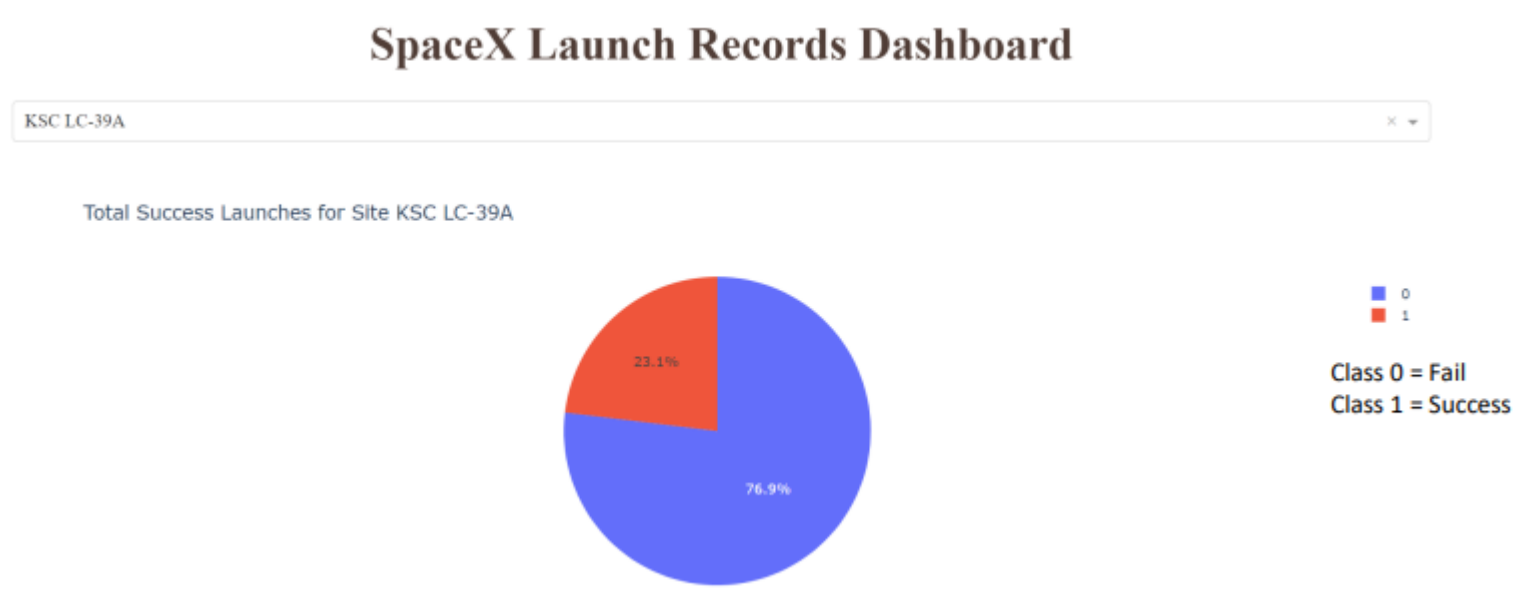
Plotly Dash – Launch Success by Site

- The purple portion represent KSC LC-39A. This site had the highest number of Total Successful Launches.



Plotly Dash – # of Launches by Site

- Launch Site KSC LC-39A had a 76.9% success rate. This was the highest success rate among all launch sites for Falcon 9 Rockets.



Plotly Dash – Payload vs Launch Outcome

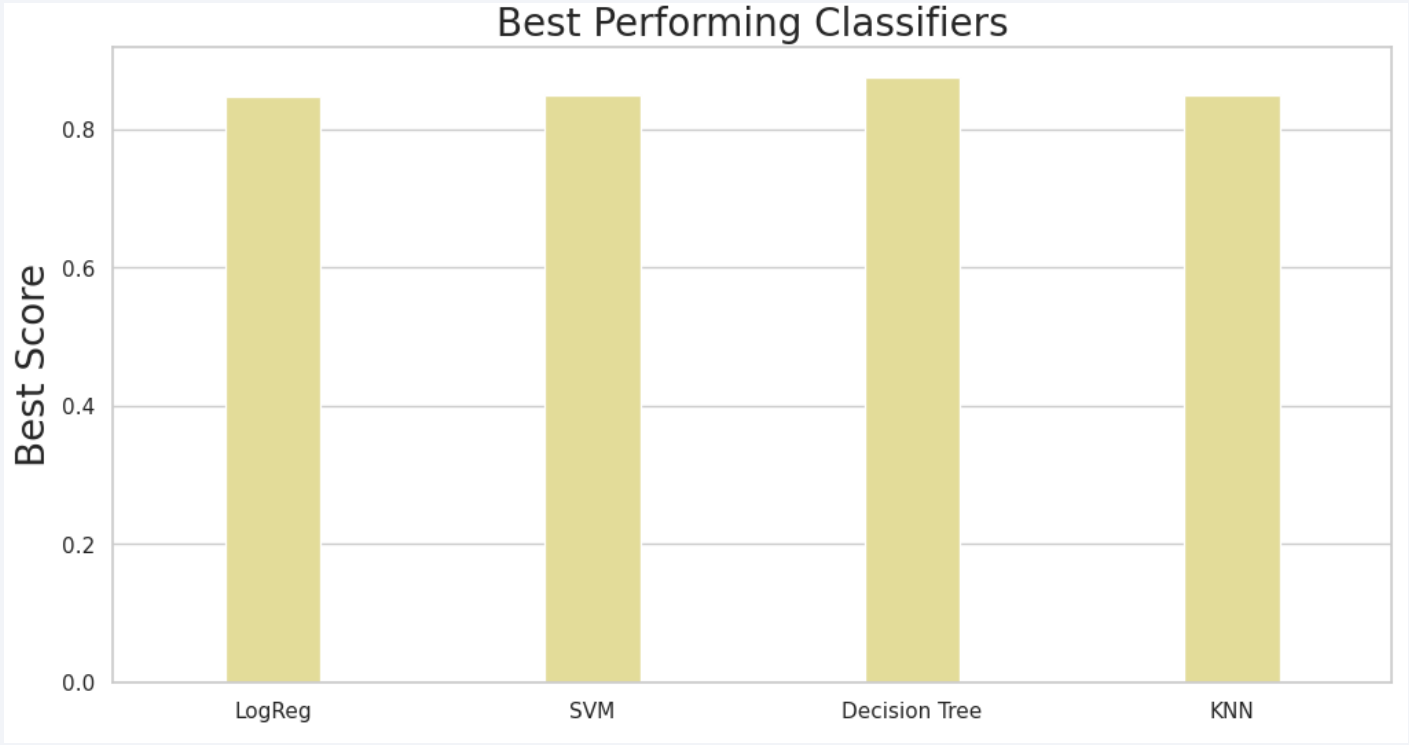
- Payloads between 2k and 5k kgs had the highest rate of success.

Section 5

Predictive Analysis (Classification)

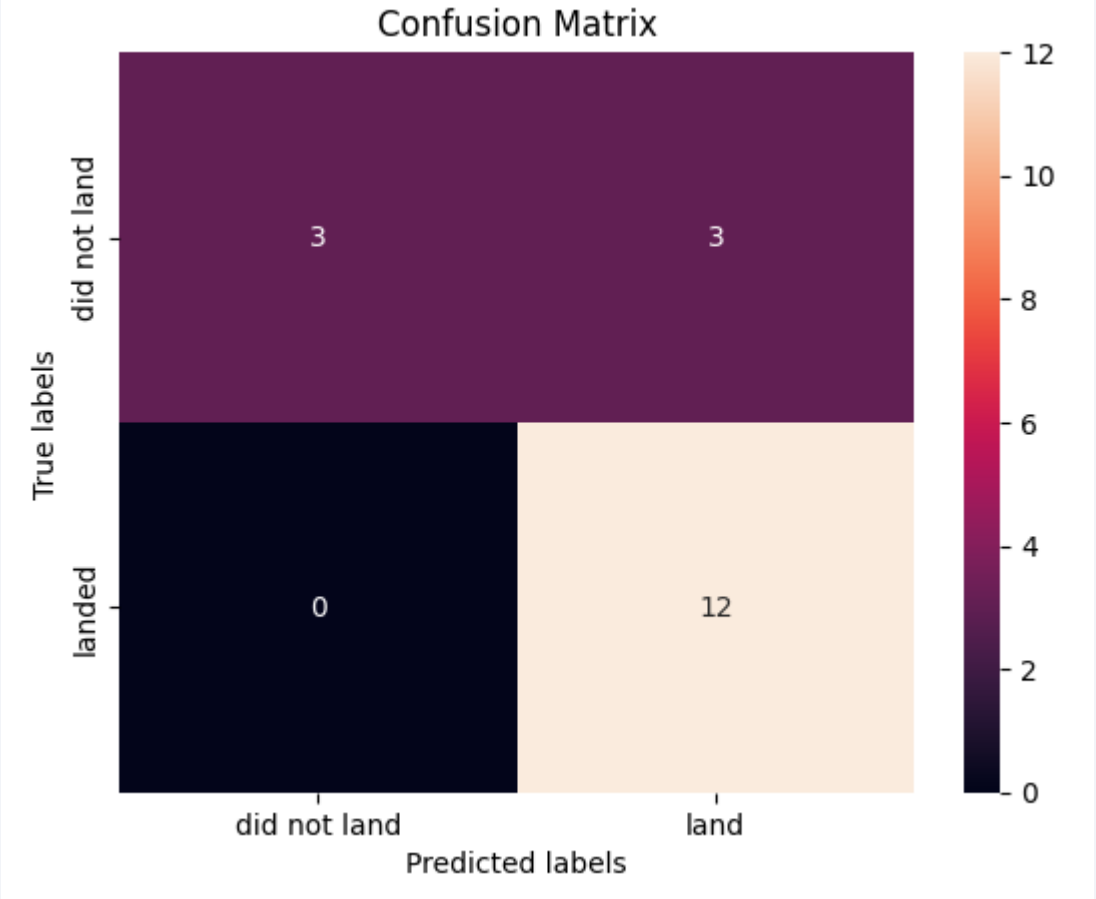
Classification Accuracy

- This bar chart displays the Accuracy score for each model type.
- The Decision Tree model has the highest score of 0.875.



Decision Tree - Confusion Matrix

- The confusion matrix displays information regarding True Positives, False Positives, True Negatives, and False Negatives.
- This model had 3 False Positives.
- Precision = TP / (TP + FP)
  - 12 / 15 = .80
- Recall = TP / (TP + FN)
  - 12 / 12 = 1
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
  - 2 * (.8 * 1) / (.8 + 1) = .89
- Accuracy = (TP + TN) / (TP + TN + FP + FN) = .833



Conclusions

- Model Performance
  - All models performed similary well but the Decision Tree slightly edged out LogReg, SVM, and KNN by having a higher Accuracy score.
- Launch Site Location
  - All launch sites, as evidenced by the Folium Mapping, are placed in the southern Regions of the country along the coast.
- Launch Success
  - KSC LC-39A Launch site had the highest success rate at 76.9%.
- Orbit Success
  - ES-L1, GEO, and HEO had a 100% success rate for Launches.
- Payload Mass
  - In general, the higher the payload mass the more likely the launch was successful.

Thank you!