Winning Space Race with Data Science

Garret

6 October 2024

Outline

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

Executive Summary

- Methodologies used in this data analysis process are as follows :
  1. Data Collection
  2. Data Wrangling
  3. Exploratory Data Analysis using SQL
  4. Exploratory Data Analysis using Pandas and Matplotlib
  5. Interactive Visual Analytics with Folium
  6. Building an Interactive Dashboard with Plotly Dash
  7. Classification (with Predictive Analysis)

- Summary of all results :
  - By visualizing the findings from EDA, features that would be the best predictors for a launch's success are found.
  - Decision Tree has the highest accuracy compared to decision tree, svm and knn- with an accuracy of 94.44% in predicting the success of a launch.

Introduction

- One of the advantages that SpaceX has compared to other companies in the same field is the fact that the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Unlike other rocket providers, SpaceX's Falcon 9 Can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash as shown in this clip. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

- Desirable answers:

  - The price of each launch.
  - Determine if SpaceX will reuse the first stage of the rocket and analyze contributing factors.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX's public API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Performed pre-processing, handling missing values
  - 'Class' feature engineering, a binary value indicating whether or not a landing was successful.

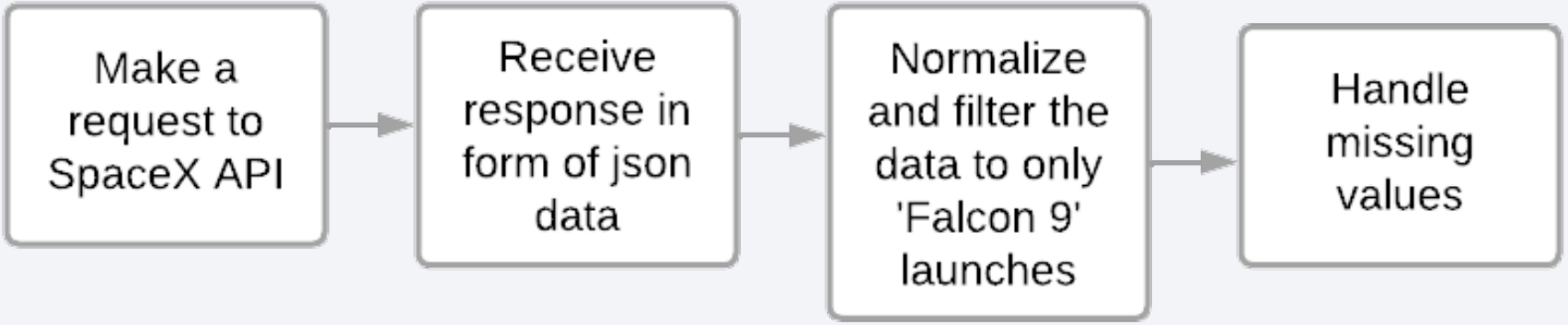
Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Normalized Data
  - Split the data into training and testing datasets
  - Used 4 different machine learning models to find the best performance by utilizing automatic hyperparameter tuning (GridSearchCV)

Data Collection – SpaceX API

- The first step of data collection uses SpaceX's public API listed here : (https://api.spacexdata.com/v4/rockets/)
- The flow of the process is shown below :



Source Code:
https://github.com/GarretJT/SpaceXCapstone/blob/main/1.%20Collecting%20Data/jupyter-labs-spacex-data-collection-api.ipynb

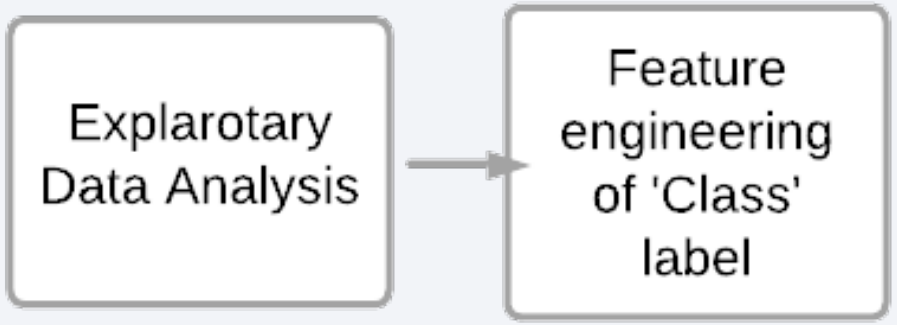
Data Collection – WebScraping

- The second step of data collection uses WebScraping on SpaceX's Wikipedia page listed here: (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- The flow of the process is shown below :

Source Code:
https://github.com/GarretJT/SpaceXCapstone/blob/main/1.%20Collecting%20Data/jupyter-labs-webscraping.ipynb

Data Wrangling

- Feature engineering of the 'Class' label was done, by creating a variable 'landing outcome' from the 'Outcome' column.
- 'Class' has a binary value (0 or 1). If the value is zero, the first stage did not land successfully; one means the first stage landed successfully.



Source Code:
https://github.com/GarretJT/SpaceXCapstone/blob/main/2.%20Data%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Matplotlib and Seaborn library was used to construct plots for visualization.
  The plots that is drawn are as follows:
  - Scatter plot
    Used to visualize the relationship between 2 variables, grouped by 'Class' like PayloadMass vs Flight Number.
  - Bar plot
    Used to visualize the success rate (of landing) by orbit type.
  - Line plot
    Used to visualize the yearly trend of launch success.

Source Code:
https://github.com/GarretJT/SpaceXCapstone/blob/main/4.%20EDA%20Visualization/edadataviz.ipynb

EDA with SQL

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in
- year 2015; and
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Source Code:
https://github.com/GarretJT/SpaceXCapstone/blob/main/3.%20EDA%20with%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers represent specific locations, such as launch sites.
- Circles highlight areas surrounding particular coordinates, like the NASA Johnson Space Center.
- Marker clusters group multiple events occurring at the same location, such as multiple launches at a single site.
- Lines is used to show the distance between two coordinates.

Source Code:
https://github.com/GarretJT/SpaceXCapstone/blob/main/5.%20Interactive%20Data%20Visualization/lab_jupyter_launch_site_location.ipynb

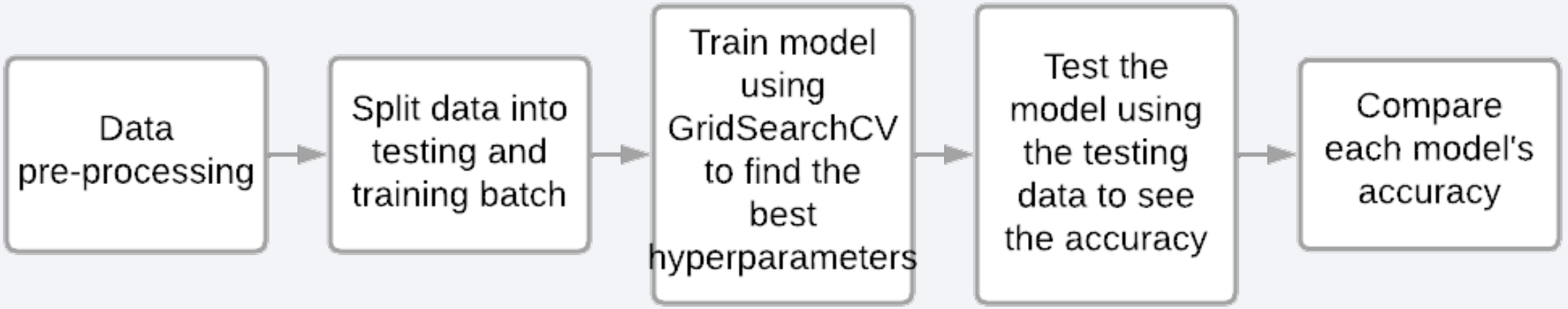
Build a Dashboard with Plotly Dash

- The dashboard contains a pie chart displaying the percentage of successful launches per launch site, which can be selected by the user.
- Also contains a payload range slider which can be selected by the user as a filter. This is done to create a user-friendly visualization dashboard to identify useful insight such as which launch site has the best success rate given a set range of payload mass.

Source Code:
https://github.com/GarretJT/SpaceXCapstone/blob/main/5.%20Interactive%20Data%20Visualization/dash/spacex_dash_app.py

Predictive Analysis (Classification)

- 4 machine learning models were used (KNN, Logistic regression, SVM, Decision Tree) in order to find the best performing model for predicting the success of a launch.
- To find the best set of hyperparameters for each model, GridSearchCV was used.



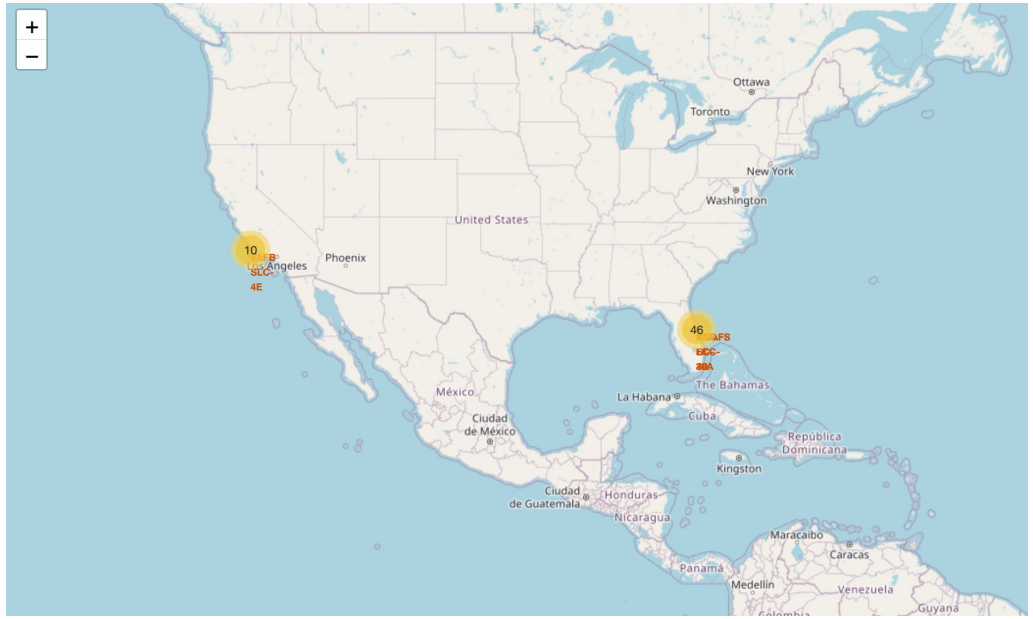
Results - EDA

- Exploratory data analysis results:

  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of ‘F9 v1.1’ booster is 2,928 kg;
  - The orbits SSO, ES-L1, GEO and HEO has a 100% success rate
  - The first landing outcome that is successful happened in 2015;
  - The overwhelming majority of the mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The average success rate of launch has an upward trend that increases over time.

Results – Interactive Visualization (Folium)

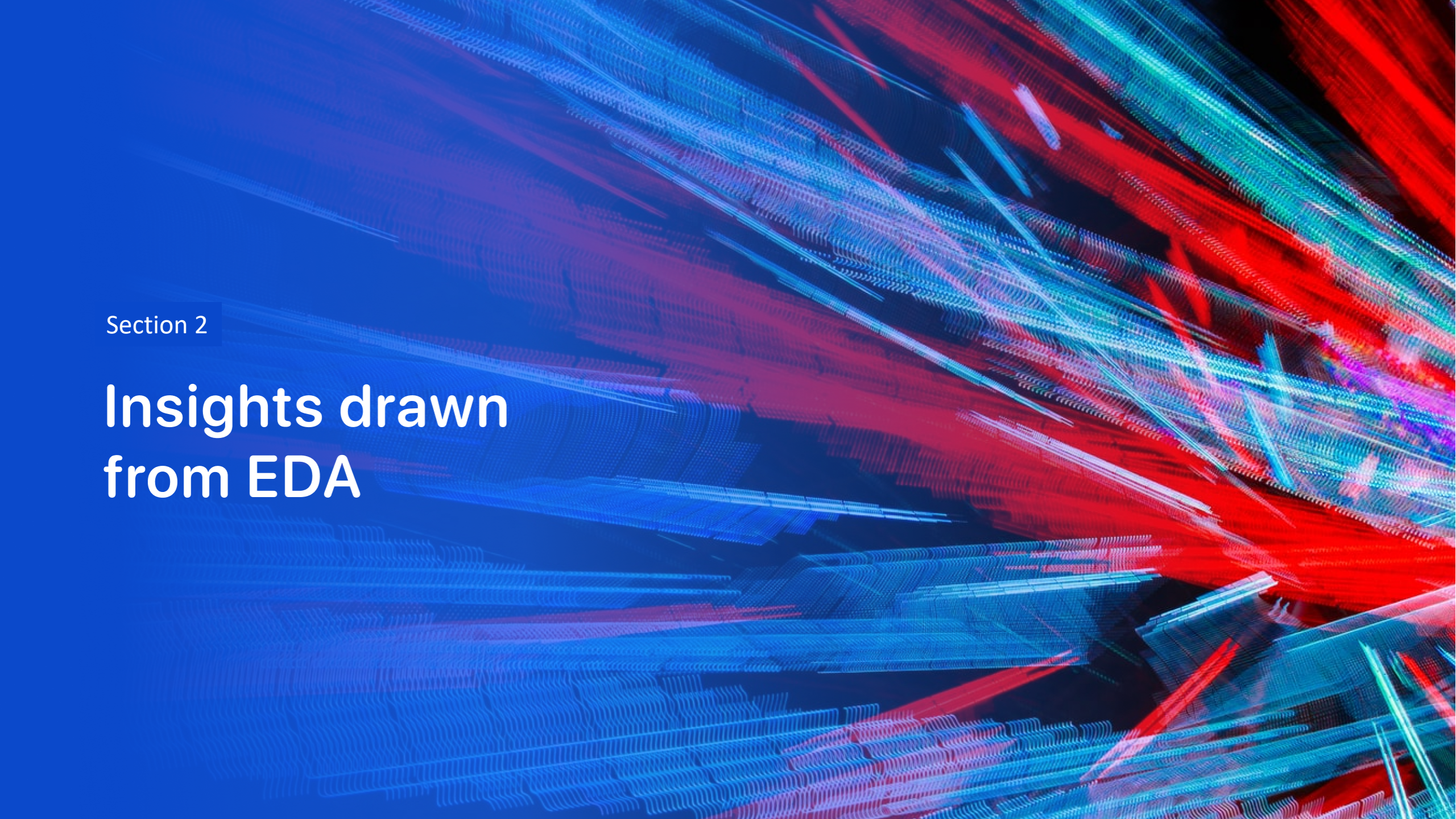
- Interactive analytics demo in screenshots
  From the visualization done through folium map, we can deduct that the majority (46) of launches are done in the east coast as opposed to the west coast.
  We can also see that all launch sites are a located the farthest away as it could be from the mainland (the coasts).



Results – Predictive Analysis Results

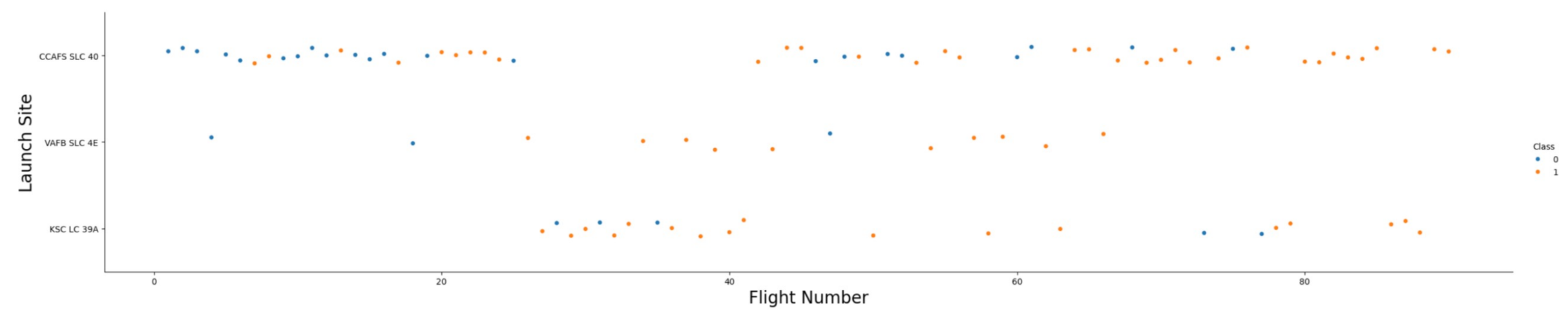
- After evaluating the 4 models, it is shown that Decision Tree Classifier has the best performance, with an accuracy of 94.44% on the testing data.

Machine Learning Model	Test Accuracy
KNN	83.34%
LogReg	83.34%
Decision Tree	94.44%
SVM	83.34%


Section 2
Insights drawn from EDA

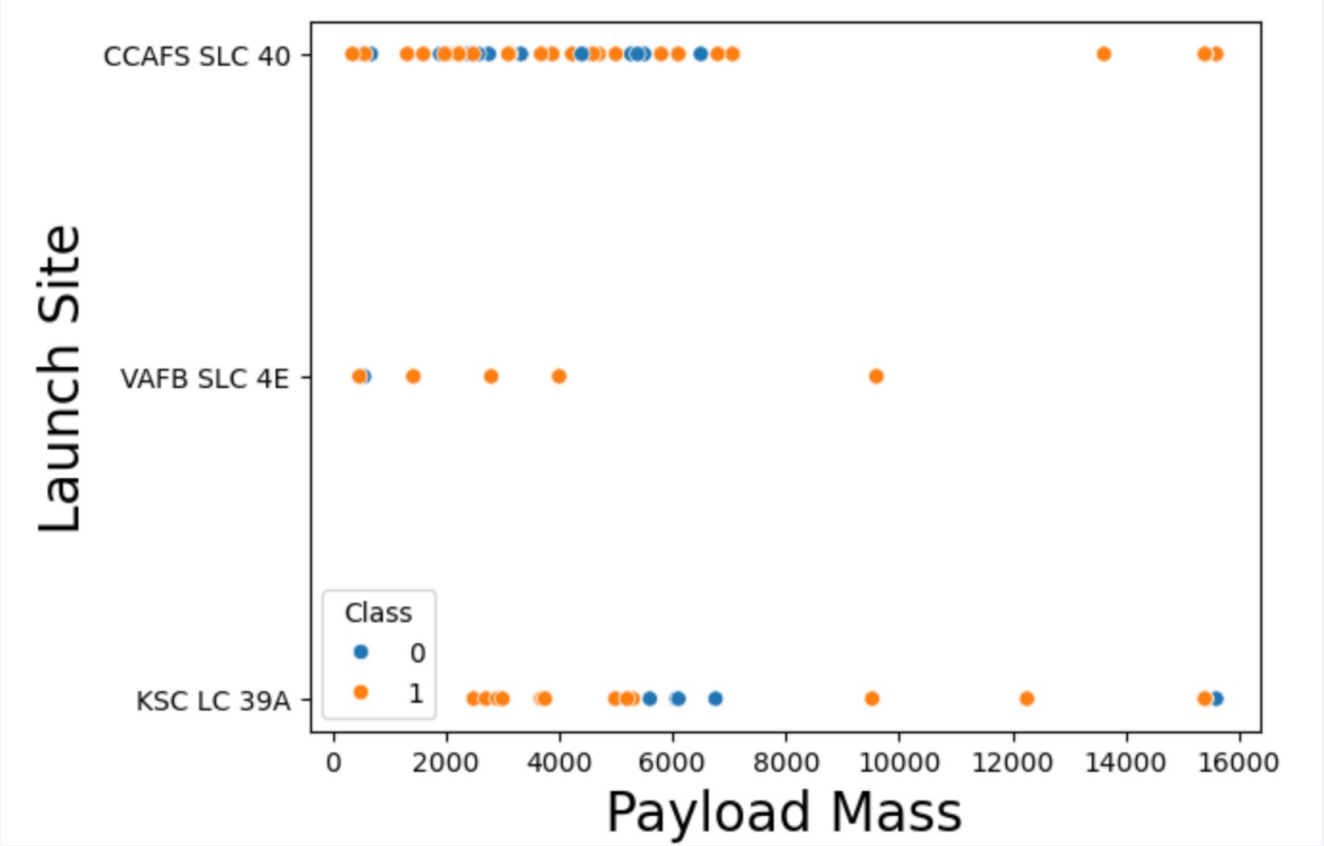
Flight Number vs. Launch Site

- CCAFS SLC 40 has the most launches, with it being the launch site with the most diverse across flight numbers. The amount of total successful launches is similar to the total of unsuccessful launches. However, with increasing flight numbers, the success rate also increases.
- Whereas VAFB SLC 4E and KSC LC 39A has a better success to unsuccessful launch ratio, however with a lower amount of total flight number



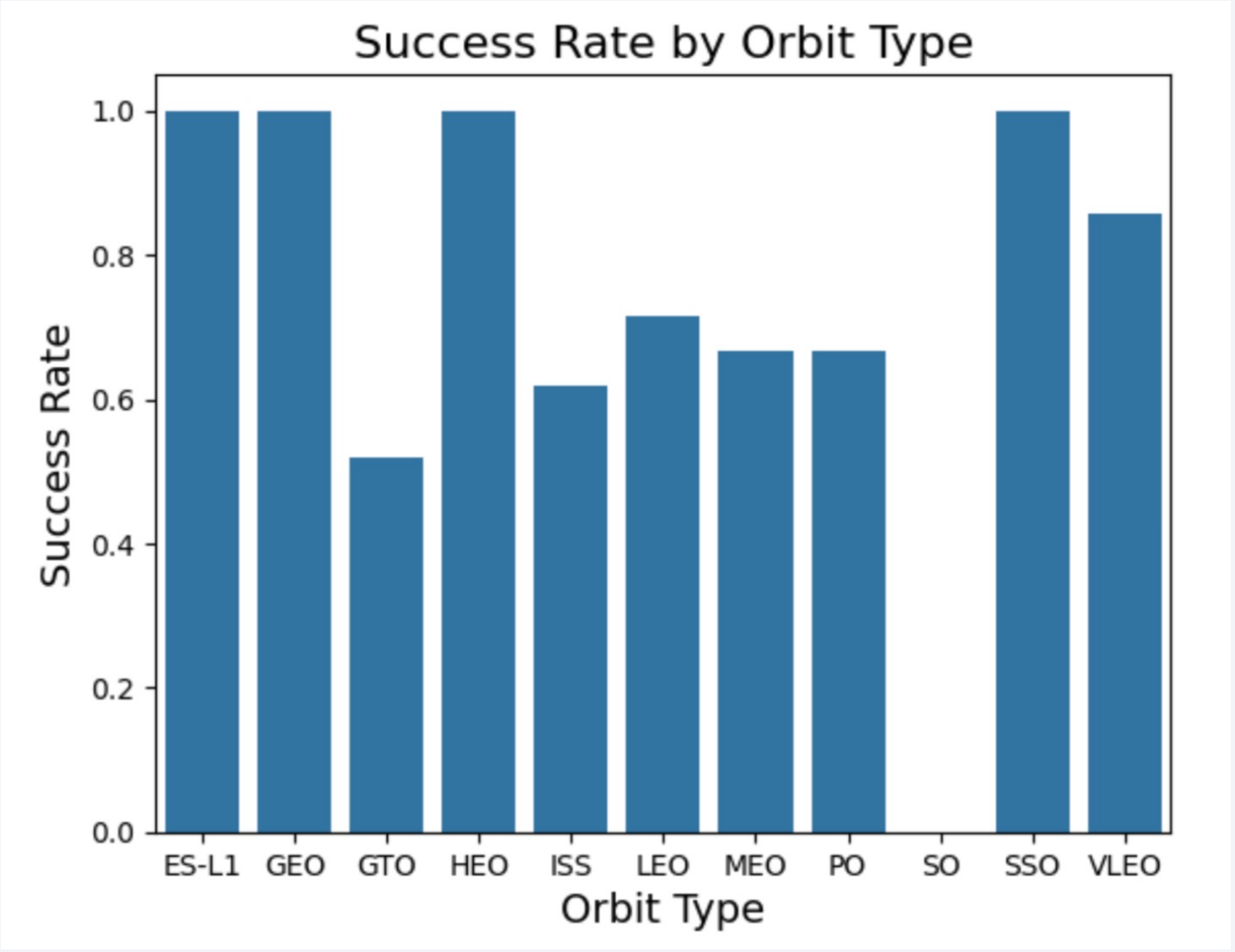
Payload vs. Launch Site

- VAFB SLC 4E has no launches with a payload mass above 10000 (which are classified as heavy-load launches)
- Most heavy-load launches (launches with payload mass above 10000) has a high success rate, with only 1 of the launches being unsuccessful



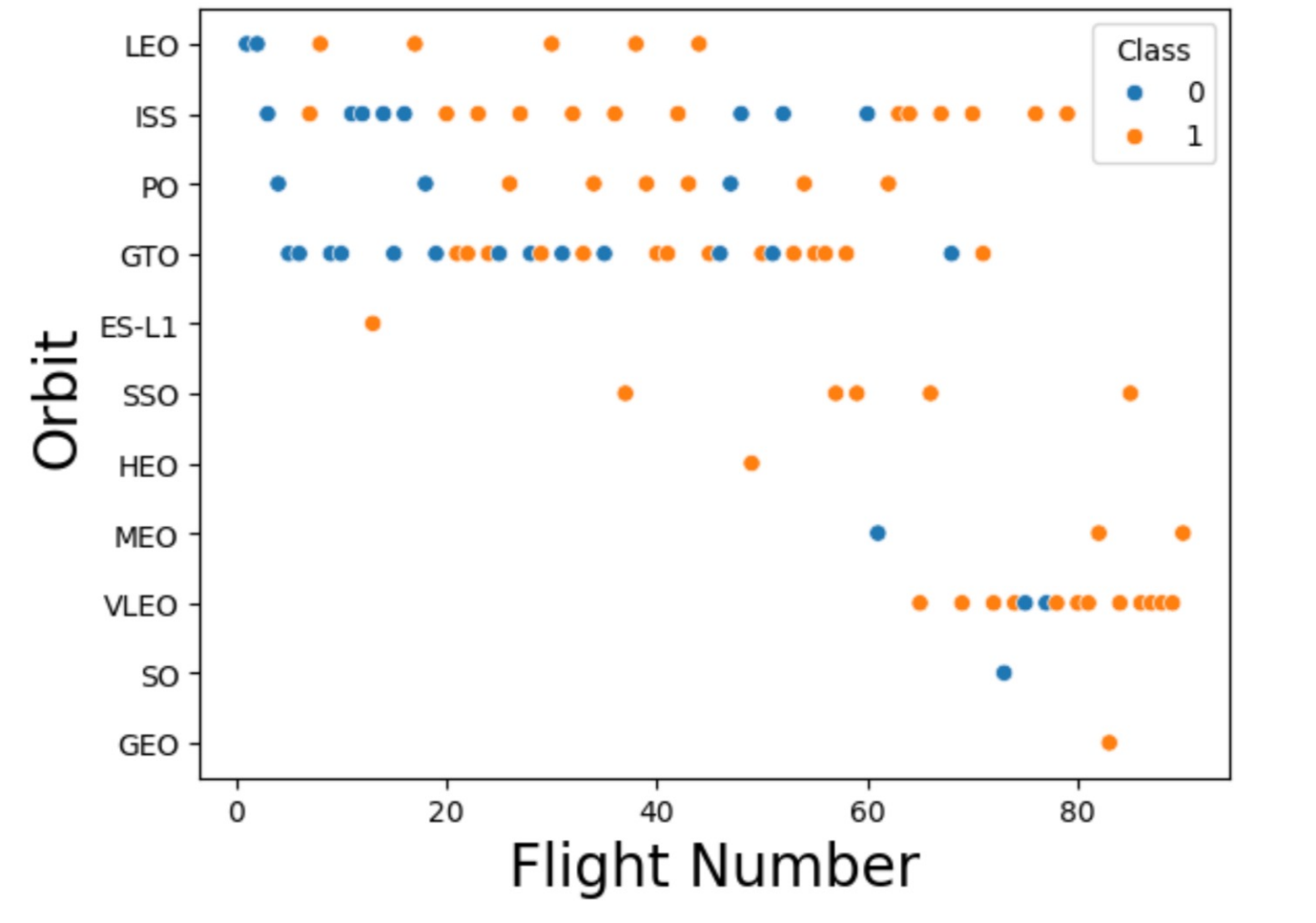
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO orbits has a 100% success rate.
- SO seems to have a 0% success rate. The dataset only has 1 occurrence of a launch with the orbit type ‘SO”, which is unsuccessful
- The orbit type with the lowest success rate following SO is GTO, with a roughly 50% success rate, followed by ISS.



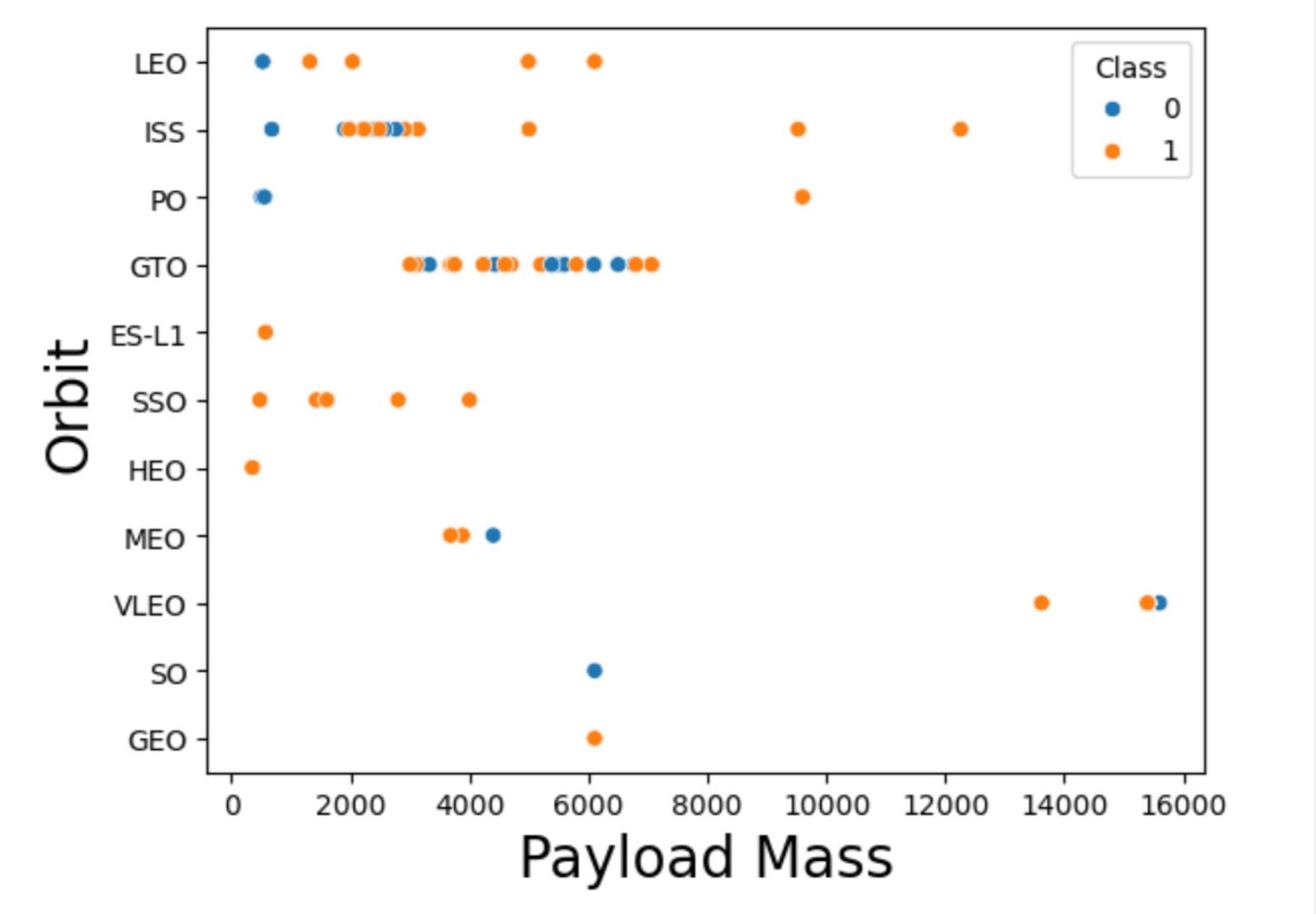
Flight Number vs. Orbit Type

- HEO, ES-L1, GEO and SSO has a 100% success rate across flight numbers.
- Across flight numbers, for all orbits, there seems to be an upward trend of increasing success launches



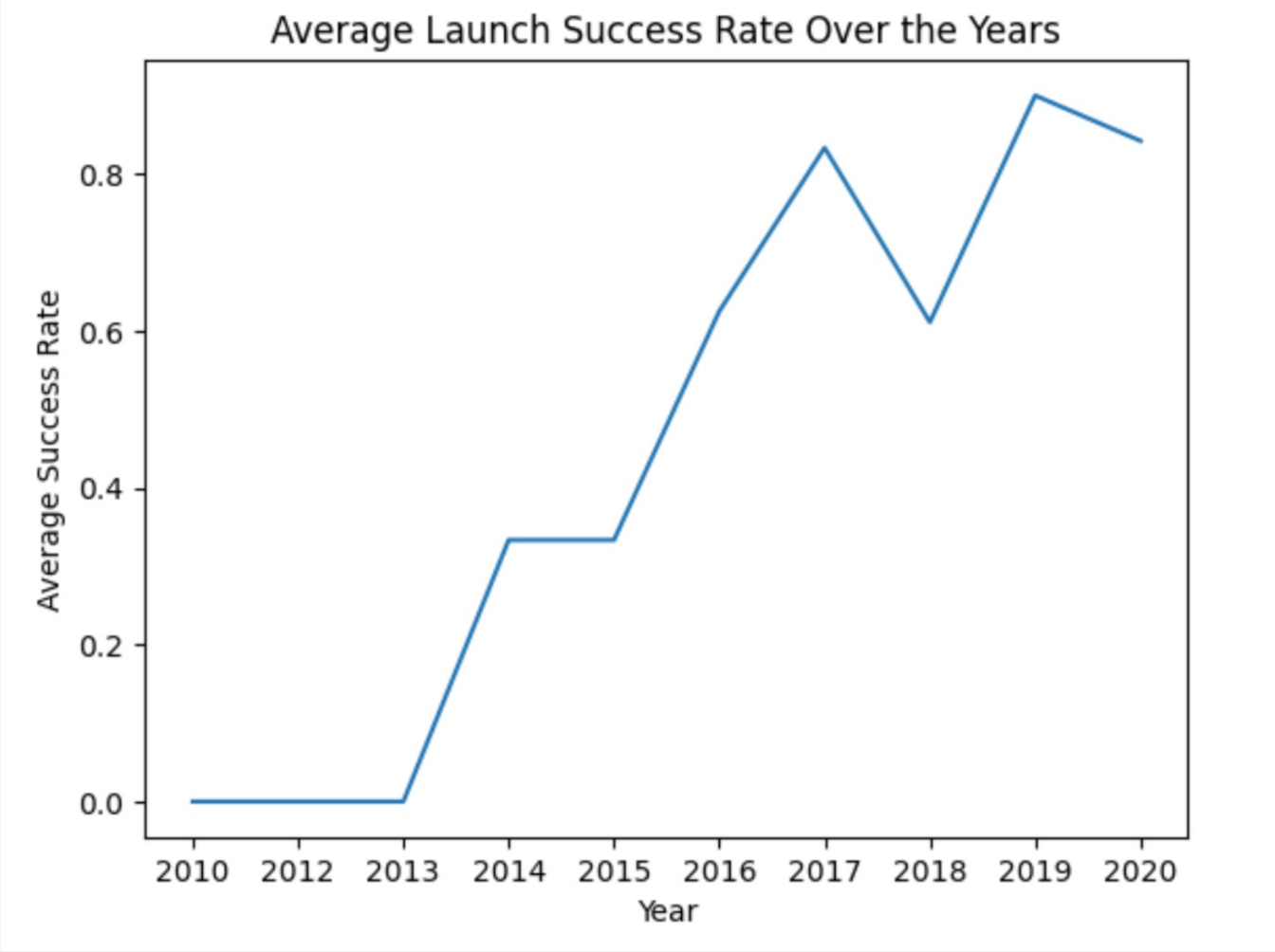
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- The average success rate has a positive correlation with time, meaning that it increases as time goes by.
- 2 noticeable increase in average success rate is on 2013 and 2015.



All Launch Site Names

- SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
- Distinct is used to find unique entries (to ensure that no duplicates are returned)

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

Launch Site Names Begin with 'CCA'

- SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
- The LIKE function is used to find Launch Site names that begin with CCA, with a % at the end to indicate that there are no constraints for what the following characters are after CCA. Limit 5 is used to display the first 5 records

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

- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE '%NASA%';
- SUM is used to return the total payload mass, where the customer has 'NASA' in it, without constraints for the characters before and after by using %

SUM(PAYLOAD_MASS__KG_)
107010

Average Payload Mass by F9 v1.1

- %%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';
- AVG is used to find the average payload mass for Booster_Versions that begin with F9 v1.1, using % at the end to include the F9 v1.1 booster children (such as F9 v1.1 B1011)

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- %sql SELECT Date FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "Success (ground pad)" ORDER BY Date ASC LIMIT 1
- WHERE "Landing_Outcome" LIKE "Success (ground pad)" is used to only return the landing outcomes that are successful (for ground pad).
- Using ORDER BY Date ASC to put the oldest date record on top, then LIMIT 1 to only select the first record

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
- WHERE "Landing_Outcome" LIKE 'Success (drone ship)' filters to only record with landing outcome is 'Success (drone ship)',
- Uses AND to include only payload mass that are between 4000-6000 on top of the previous WHERE filter

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

Total Number of Successful and Failure Mission Outcomes

- Uses UNION ALL to combine the results of both queries (for Success and Failure).
- Uses ‘AS’ to name the table header for clarity
- Uses WHERE “Mission_Outcome” LIKE to filter mission outcomes that are Success, and also Failure.

Outcome	Count
Success	100
Failure	1

```
%%sql SELECT 'Success' AS Outcome, COUNT("Mission_Outcome") AS Count
FROM SPACEXTABLE
WHERE "Mission_Outcome" LIKE 'Success%'
UNION ALL
SELECT 'Failure' AS Outcome, COUNT("Mission_Outcome") AS Count
FROM SPACEXTABLE
WHERE "Mission_Outcome" LIKE '%Failure%';
```

Boosters Carried Maximum Payload

- Uses a subquery to find the MAX amount of payload mass across all records, then uses the result of that for a WHERE filter, to find Booster Versions that holds the max amount of payload across all records

```
%%sql SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
```

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

- Finds the Month [using substr("Date", 6,2)], landing outcome, booster version and launch site
- Where the landing outcome is a failure, and in the year 2015 using a similar subtr method.

```
%%sql SELECT
    substr("Date", 6, 2) AS Month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM
    SPACEXTABLE
WHERE
    "Landing_Outcome" LIKE '%Failure%'
    AND substr("Date", 1, 4) = '2015';
```

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

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

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

- Ranks the count of landing outcomes between the dates specified
- Using GROUP BY to group the records to find the outcome count for each landing outcome entries.
- Using ORDER BY DESC to display the result in descending order.

```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY Outcome_Count DESC
```

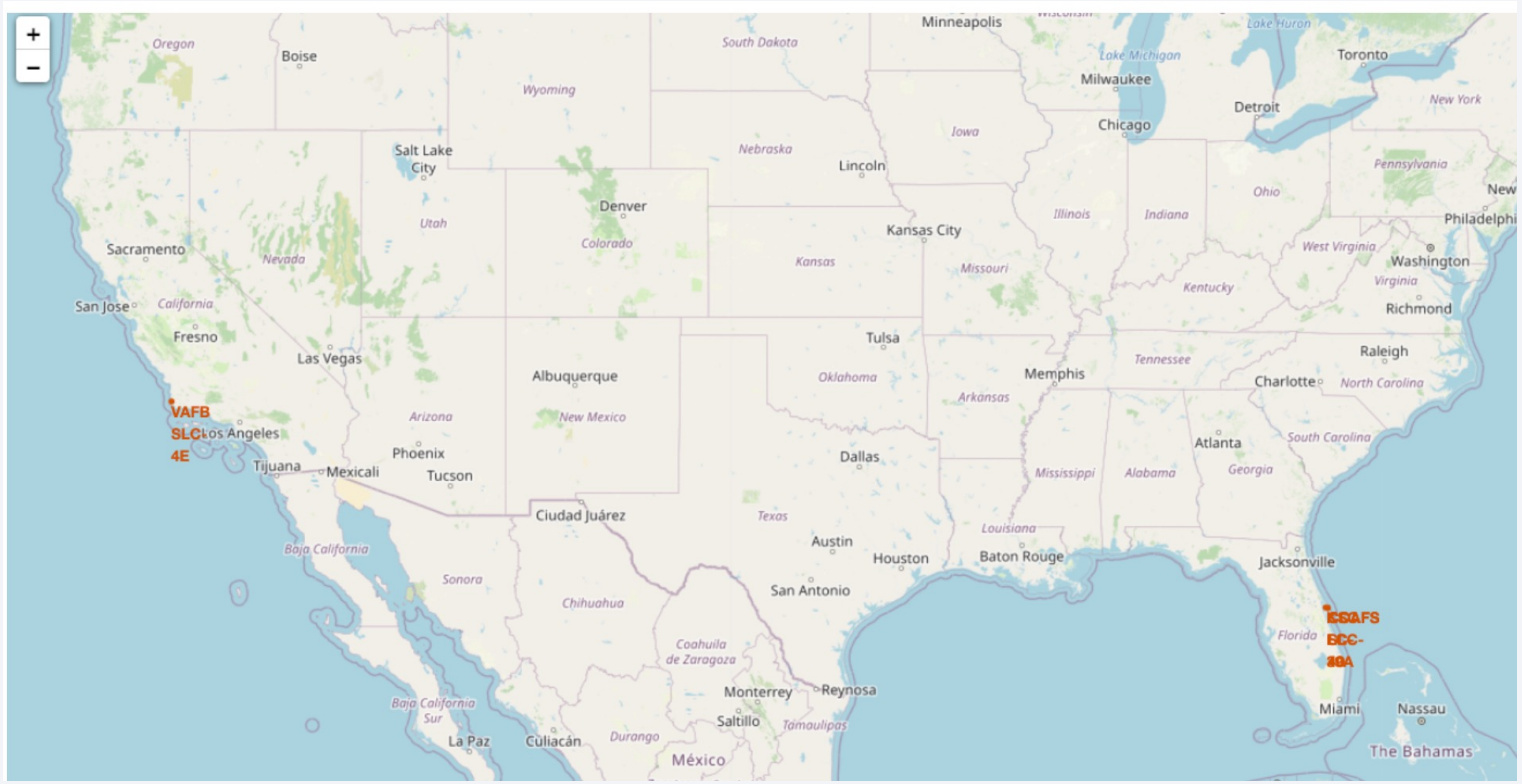
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Outcome_Count
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


Section 3
Launch Sites
Proximities Analysis

Location of all launch sites

- An interactive map display of all launch site locations. We can see a pattern that all launch sites are located in the coast, with more of them on the east coast.

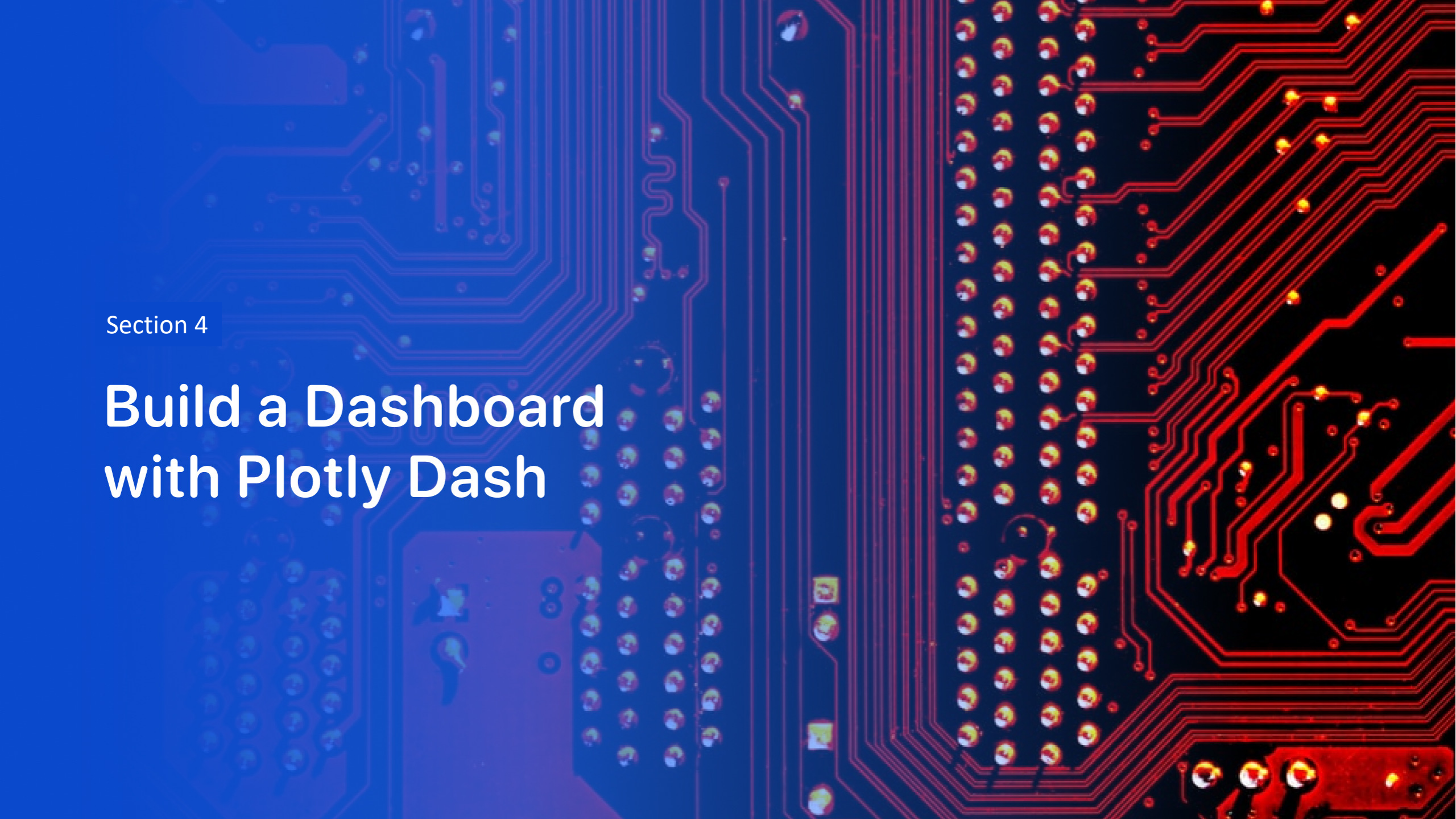


Launch Outcomes

- The launch outcome is marked with a green (for successful) and red (for unsuccessful) launches, grouped by each site. Below is an example of the CCAFS SLC-40 launch site

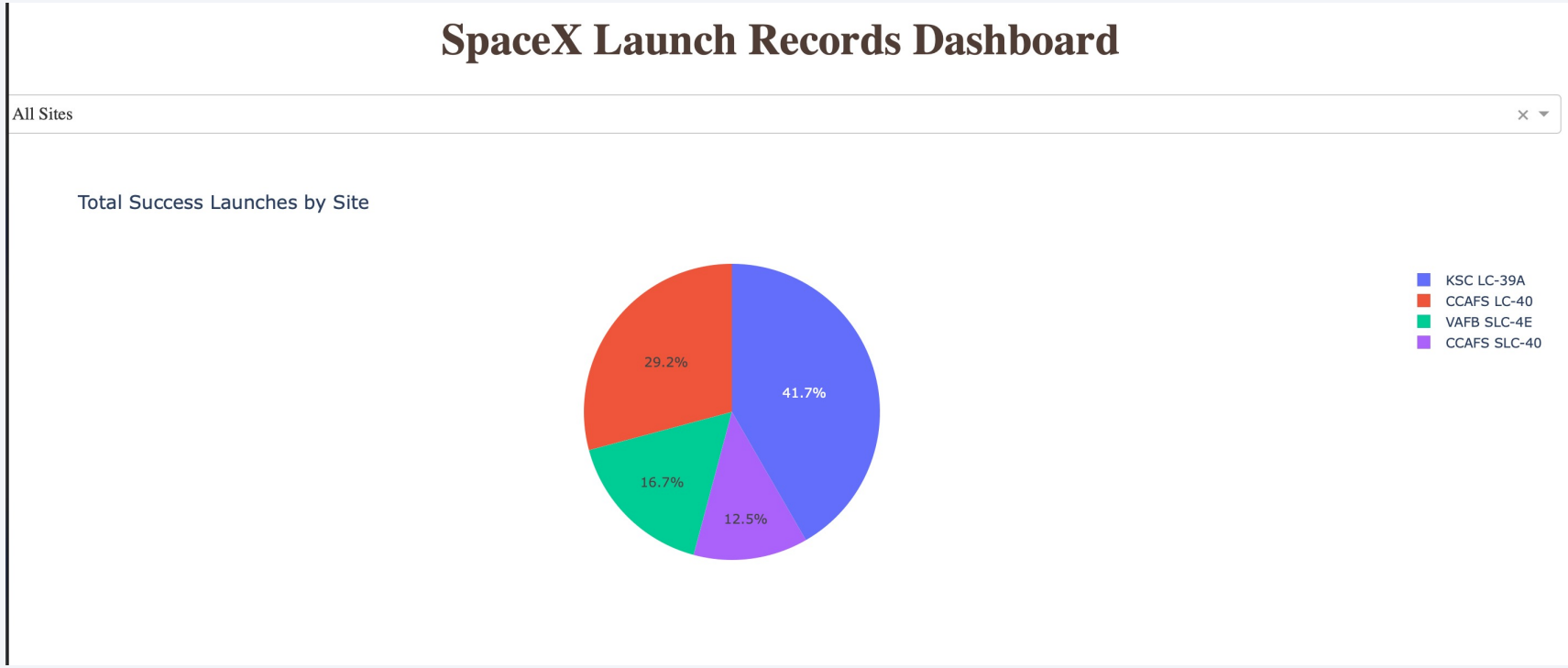
Deliberate placement of launch sites

- We can see that the KSC LC-39A launch site is located a significant distance away from civilization, as seen that it is 15.23 km away from the nearest railway. The neighboring launch sites are also located farther away, closer to the sea.


Section 4
Build a Dashboard with Plotly Dash

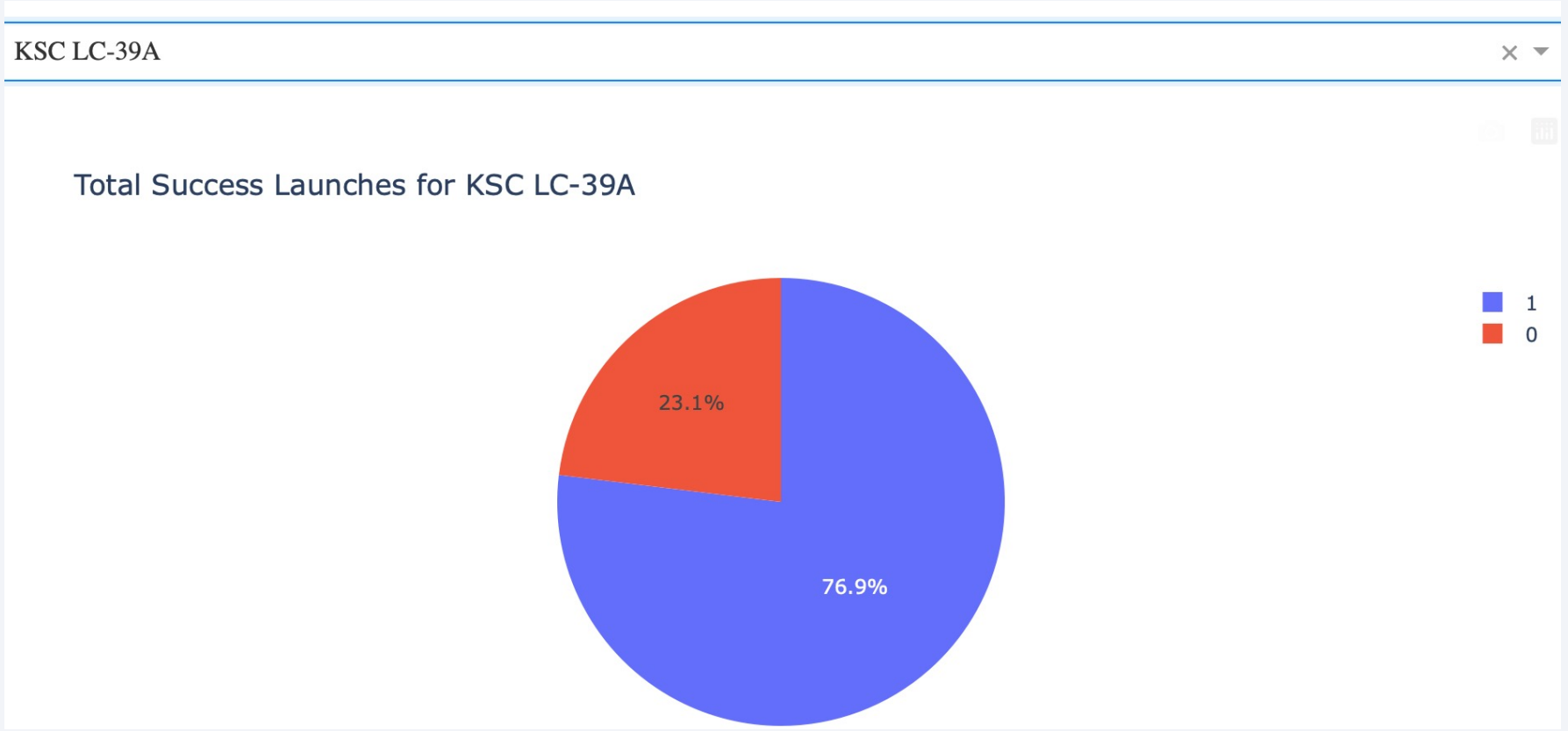
Pie Chart of Total Success Launches for All Sites

- When looking at the pie chart of total success launches for all sites, it is apparent that KSC LC-39A contributes the most (41.7%), followed by CCAFS LC-40 (29.2%) which also makes sense because both of them also has the highest number of launches in general.



Launch Success ratio for KSC LC-39A

- Over ¾ of KSC LC-39A's launches are a success (as indicated by the class label 1)



Payload vs Launch Outcome

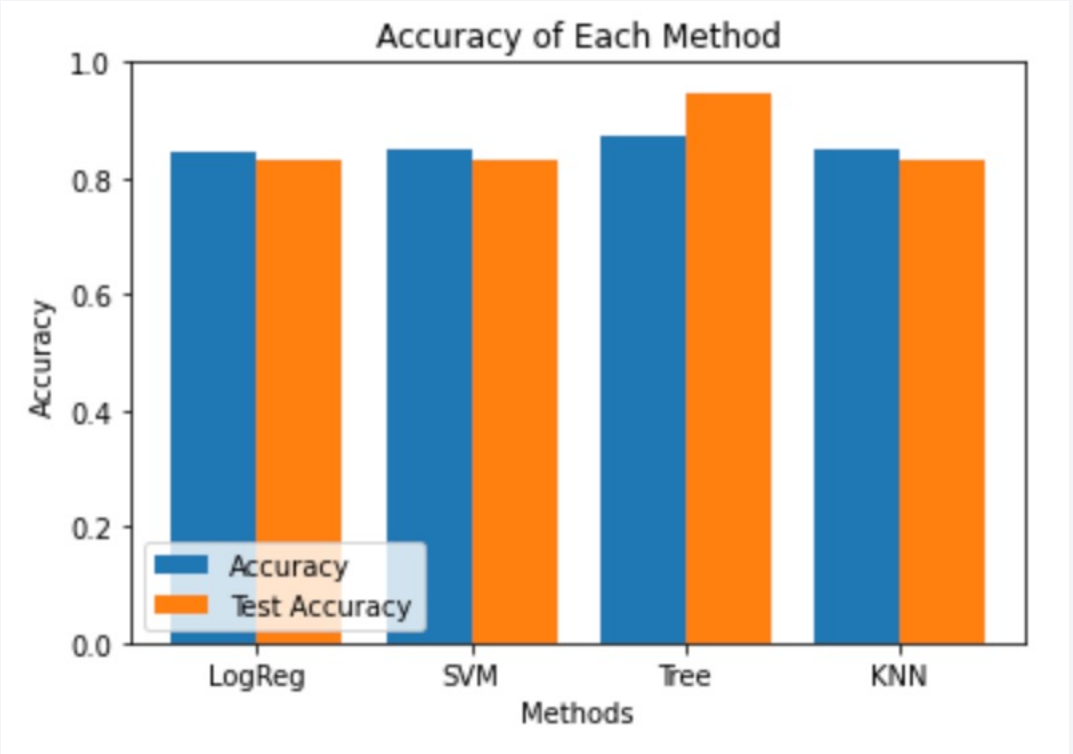
- When analyzing the correlation between payload and success for all sites, grouped by booster version category, for payloads between 0-5000 it is apparent that v1.1's majority launches are unsuccessful, while FT has a much better success ratio for the selected payload range.

Section 5

Predictive Analysis (Classification)

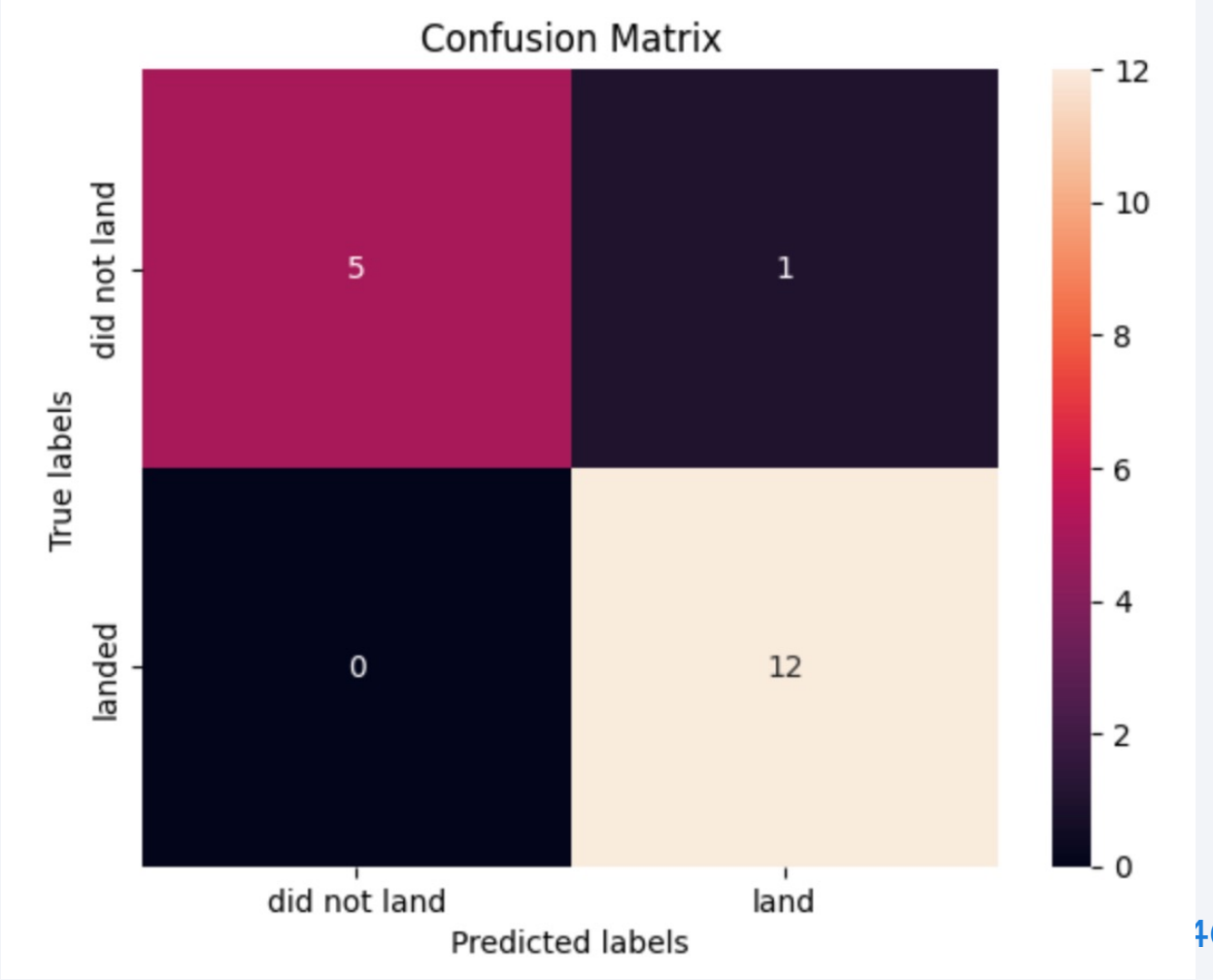
Classification Accuracy

- It is shown that Decision Tree Classifier has the best performance with an accuracy of 94.44%



Confusion Matrix

- The decision tree confusion matrix shows that it correctly classified most of the data, justifying its 94.44% accuracy
- 12 data were correctly classified as landed (True positives)
- 5 data were collecting classified as did not land (True negatives)
- And only 1 data was misclassified as land, when it was supposed to be did not land (False positive)



Conclusions

- The number of successful landing outcomes is positively correlated with time, which in theory makes sense due to the technological advancements that would be made across time.
- Most heavy load launches are successful.
- The best model to use in this case is Decision Tree Classifier, with a 94.44% accuracy for predicting the success of a launch.
- Launch sites are deliberately located in the coasts, where it is far away from civilization.

Appendix

- The main github link can be found here: https://github.com/GarretJT/SpaceXCapstone/tree/main
- All flowcharts are made with using LucidChart

Thank you!