IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hajir Al Zadjali

14/04/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- The project addresses the problem of whether SpaceX will reuse its first stage based on the landing outcome. This then helps predict the price of a launch.
- The data was collected from the SPACEX REST API as well as webscraped from tables in relevant wikis. It was then cleaned and processed. Exploratory data analysis was used to identify which features are important in determining the landing outcome.
- The data was split for training and testing purposes. Four models were built and their accuracies were compared to find the best model.
- The best model was the SVM model which was able to best classify the landing outcomes for the launches.
- It was seen that CCAFS LC-40 had the largest successful launches, whilst KSC LC-39A had the highest launch success rate.
- The booster version FT had the highest success rate, whilst v1.1 had the lowest launch success rate.

Introduction

- SpaceX Falcon 9's first stage landing is an important predictor of the cost of a launch. SpaceX advertises a launch cost of 62 million dollars, whereas its competitors provide a launch of 165 million dollars. The savings in the cost launch are attributed to SpaceX's ability to resuse the first stage.
- SpaceY would like to compete with SpaceX in the commercial space travel race. This project aims to answer the following questions:
- Will SpaceX reuse the first stage?

Section 1

Methodology

Methodology

Executive Summary

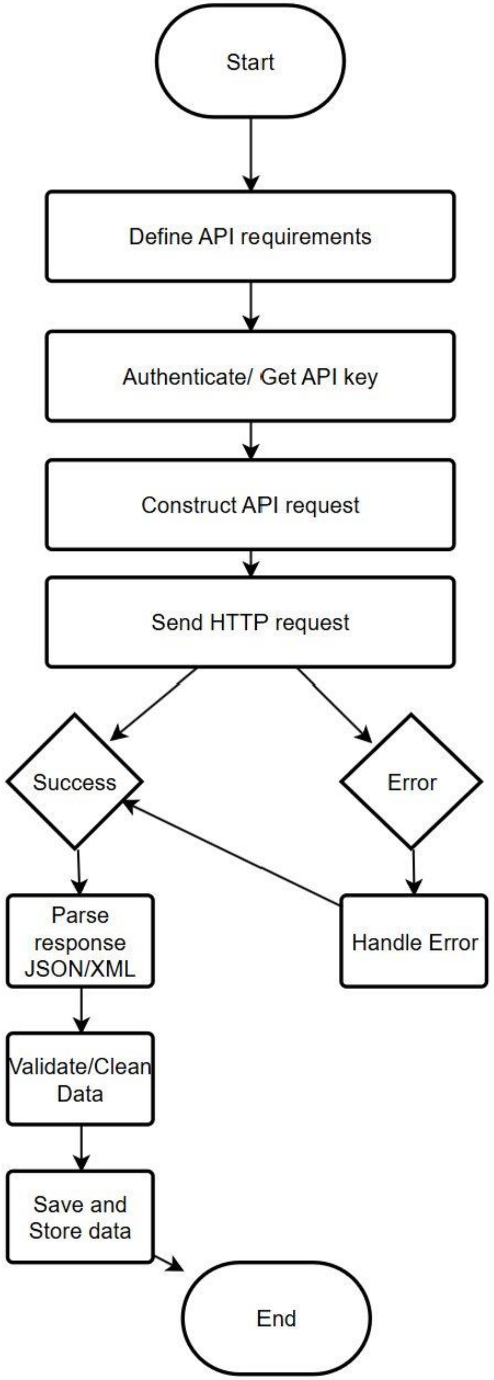
- Data collection methodology:
  - The SpaceX launch data was gathered from the SPACEX REST API, as well as from webscraping related wikipages that contain Falcon9 launch records in HTML tables.
- Perform data wrangling
  - Missing data was extracted using an API, data was also filtered to contain only Falcon 9 launch records, the null values in the PayLoadMass column were replaced by the mean, and classes were created for the Landing Outcome column.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different classification models were built and tested to find the best one.

Data Collection

- The SpaceX launch data was collected from the SPACEX REST API which contains information of the rockets used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Additionally, the Falcon9 launch data was also obtained by webscraping the relevant wiki pages that contain the launch data in HTML tables.
- Once the data is obtained it is stored in a Pandas Dataframe for further processing and analysis.

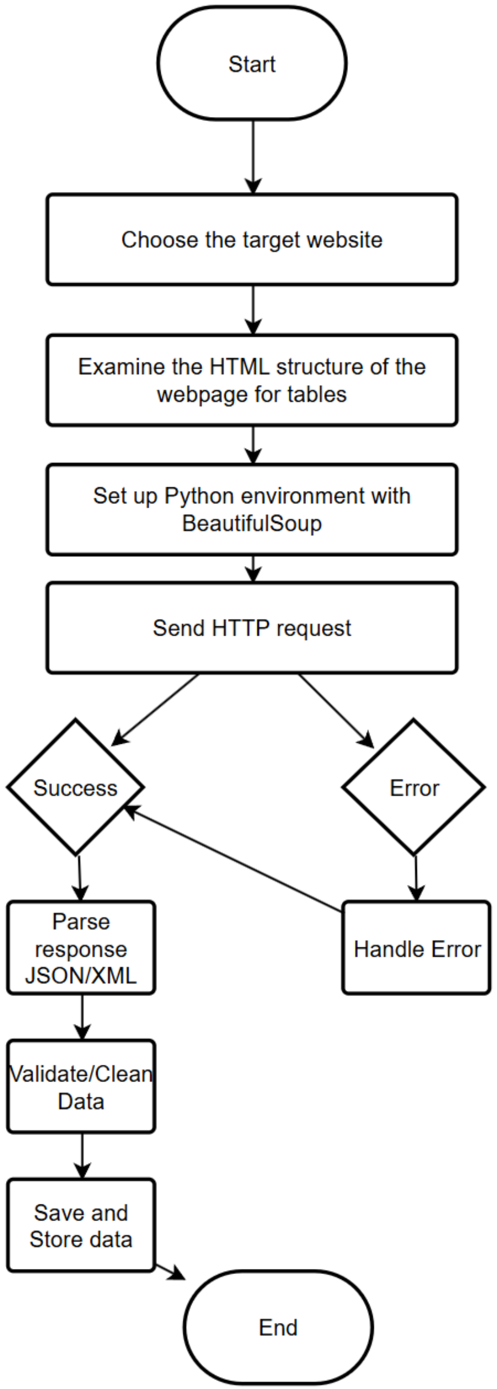
Data Collection – SpaceX API

- The SpaceX launch data was collected from the SPACEX REST API which contains information of the rockets used, payload delivered, launch specifications, landing specifications, and landing outcome.
- The URL was used to target a specific endpoint to get past the launch data. A get request was then performed to obtain the data from the API.
- This parses the data into a json file, which can then be converted into a Pandas Dataframe.
- https://github.com/hajirkhalid/Winning-the-Space-Race-with-Data-Science/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Data Collection - Scraping

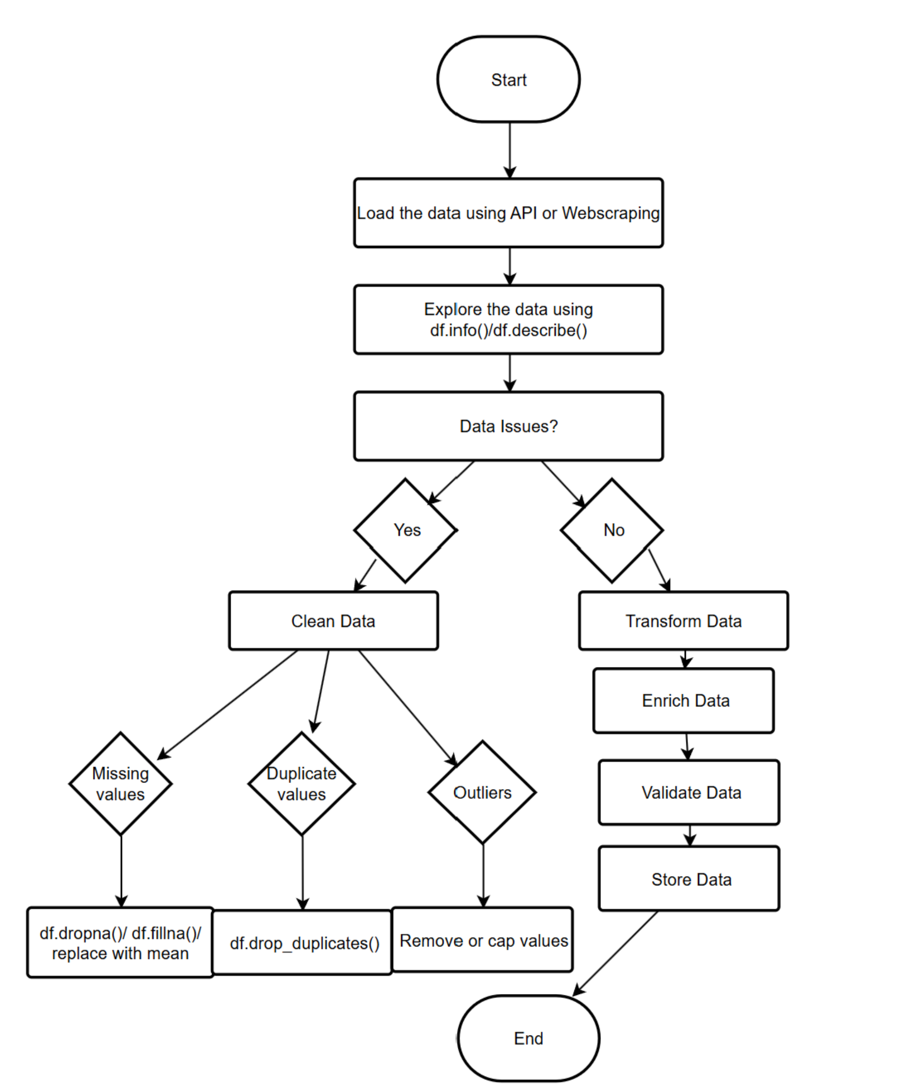
- The Python BeautifulSoup package was used to scrape some HTML tables that contain Falcon 9 launch data.
- The data was parsed from the tables and converted to a Pandas Dataframe for further analysis.
- https://github.com/hajirkhalid/Winning-the-Space-Race-with-Data-Science/blob/main/jupyter-labs-webscraping.ipynb



Start
Choose the target website
Examine the HTML structure of the webpage for tables
Set up Python environment with BeautifulSoup
Send HTTP request
Success
Error
Parse response JSON/XML
Handle Error
Validate/Clean Data
Save and Store data
End

Data Wrangling

- Some of the columns like rocket contained identification numbers instead of data. The API was used again to target different endpoints to collect data pertaining to each identification number.
- The data also contained Falcon1 booster launches which had to be filtered out.
- The null values of the PayLoadMass Column were replaced by the mean value.
- Finally, the landing outcomes were converted into classes: 0 for a bad outcome where the booster did not land, and 1 for a good outcome where the booster landed successfully.
- https://github.com/hajirkhalid/Winning-the-Space-Race-with-Data-Science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- The following plots were made:
- Scatter plots of Flight Number vs Launch Site, Pay Load Mass vs Launch Site
- Scatter plots of Flight Number vs Orbit Type, Pay Load Mass vs Orbit Type
- A bar plot of the Success Rate by Orbit Type
- A line plot of Success Rate vs Year
- A line plot was used to used to visualize the yearly trend in the success rate. The scatter plots were used to investigate any correlation between the two variables or trends and patterns between them.
- A bar plot was used since Orbit Type is categorical and the mean success rate can be visualized.
- https://github.com/hajirkhalid/Winning-the-Space-Race-with-Data-Science/blob/main/edadataviz.ipynb

EDA with SQL

- The following are some of SQL queries that were performed:
- Display the names of the unique launch sites in the space mission using DISTINCT
- Display 5 records where launch sites begin with the string 'CCA' using LIKE "CCA%" LIMIT 5
- Display the total payload mass carried by boosters launched by NASA(CRS) using SUM and WHERE
- Displayed the average payload mass carried by booster version F9 v1.1 using AVG and WHERE
- List the date of the first successful landing outcome in ground pad using MIN
- https://github.com/hajirkhalid/Winning-the-Space-Race-with-Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

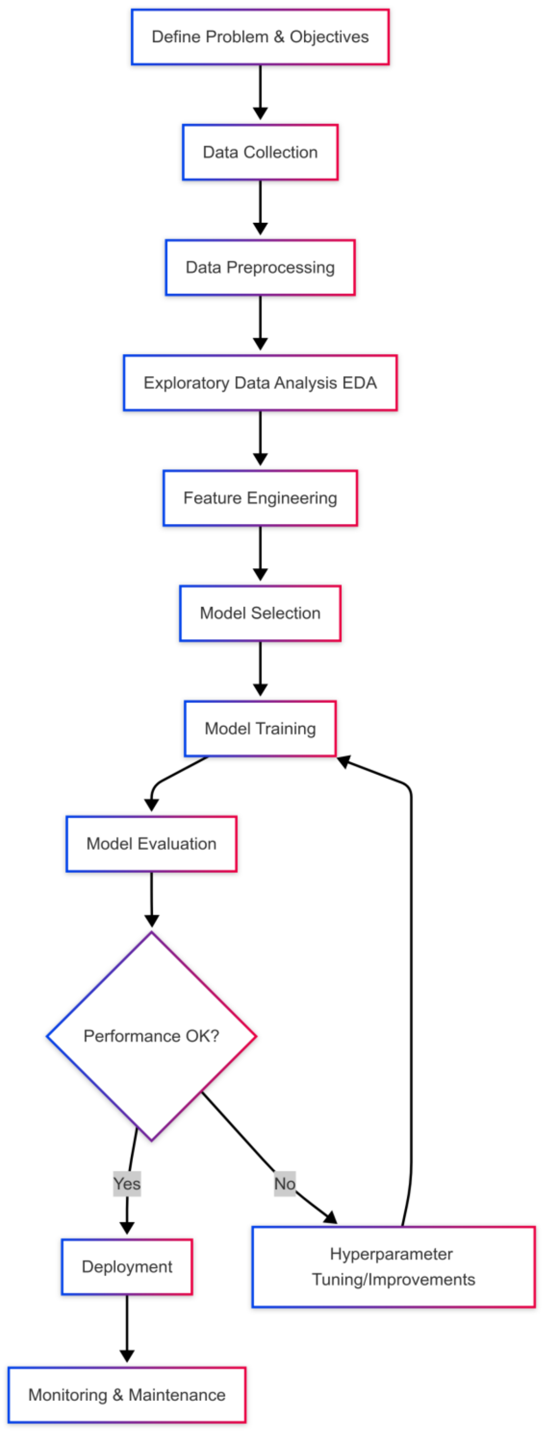
- Circles were added to the Foilum Map to mark the launch sites. Markers were added for each launch and were clustered. Color based markers were added to the marker clusters to highlight which launch sites have relatively high success rates.
- A marker was also created with distance closest to the coastline, and a line was drawn between this marker and the launch site closest to it.
- https://github.com/hajirkhalid/Winning-the-Space-Race-with-Data-Science/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A dropdown menu was added with the options to select a specific sites or all sites.
- A pie chart showing the successful launches for all sites was added, as well as individual pie charts showing the successful launches for each selected site.
- A slider was added for the Payload to be able to easily select different payload ranges and see if we can identify any patterns.
- A scatter plot of the Payload vs Launch Outcome was added to observe the correlation between the two variables. Different colors were used to mark different boosters.
- https://github.com/hajirkhalid/Winning-the-Space-Race-with-Data-Science/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

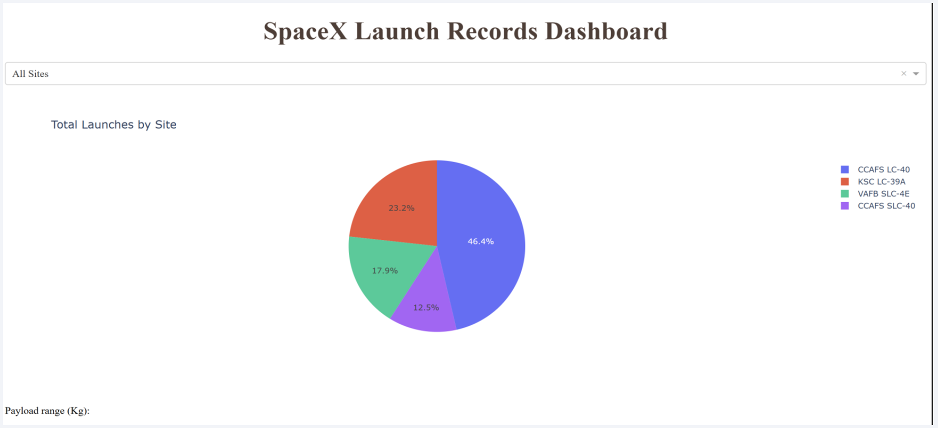
- After selecting the important features, we built a model using those features and the target variable which is the landing outcome.
- The test data was split into training and testing data after it was standardized.
- Four models were built using the training data: Logistic regression model, SVM model, decision tree classifier, and a KNN model.
- Each model was fitted to the training data and the hyperparameters were used to find the best model parameters.
- The accuracy scores and confusion matrices for the models were computed and compared to find the best model.
- https://github.com/hajirkhalid/Winning-the-Space-Race-with-Data-Science/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

- he success rate of the launches increases with time, suggesting that SpaceX will be able to reuse the first stage better as it improves its landing tehcnology.

- The highest successful launches were from site CCAFS LC-40, whilst the lowest were from VAFB SLC-4E.



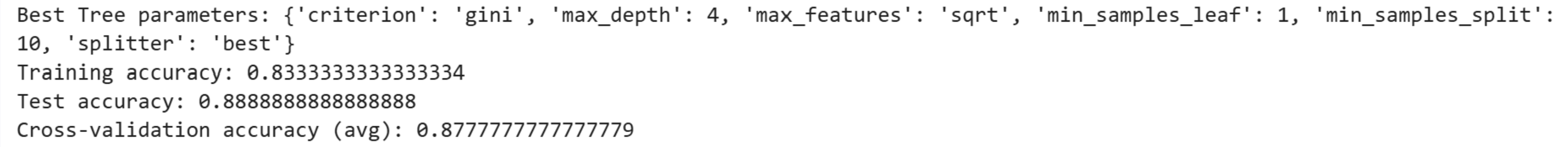
SpaceX Launch Records Dashboard

All Sites

Total Launches by Site

Payload range (Kg):

- The best model was found to be the decision tree classifier model with the following parameters:



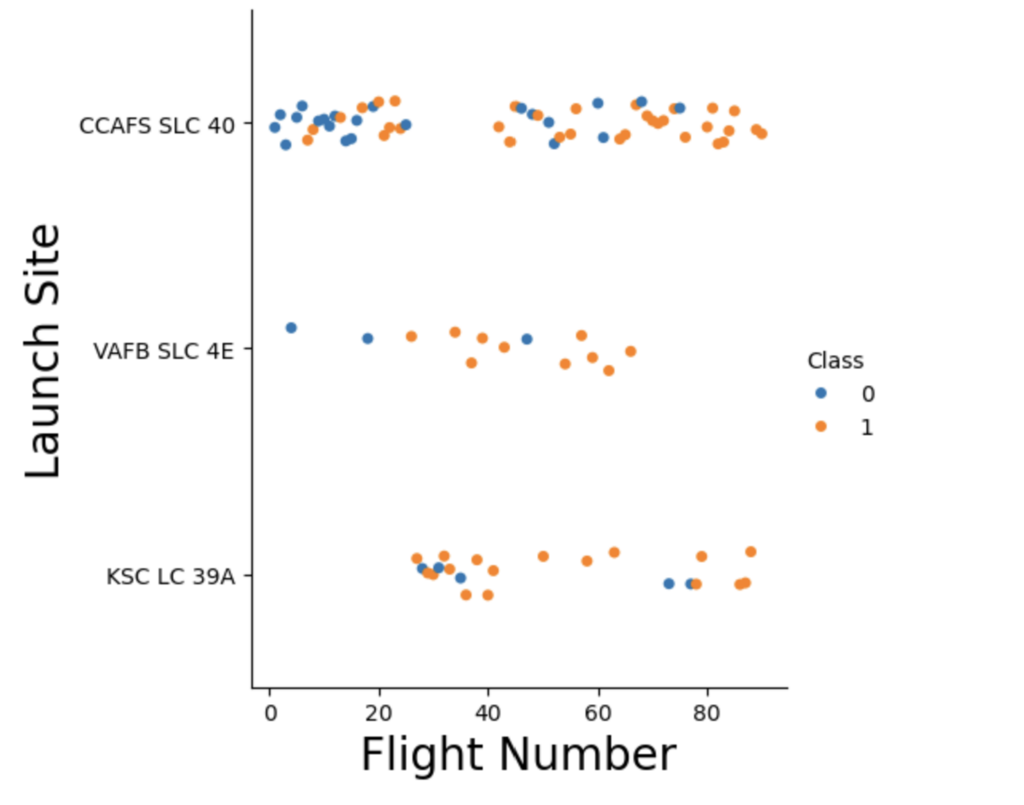
```
Best Tree parameters: {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split':
10, 'splitter': 'best'}
Training accuracy: 0.8333333333333334
Test accuracy: 0.8888888888888888
Cross-validation accuracy (avg): 0.8777777777777779
```

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

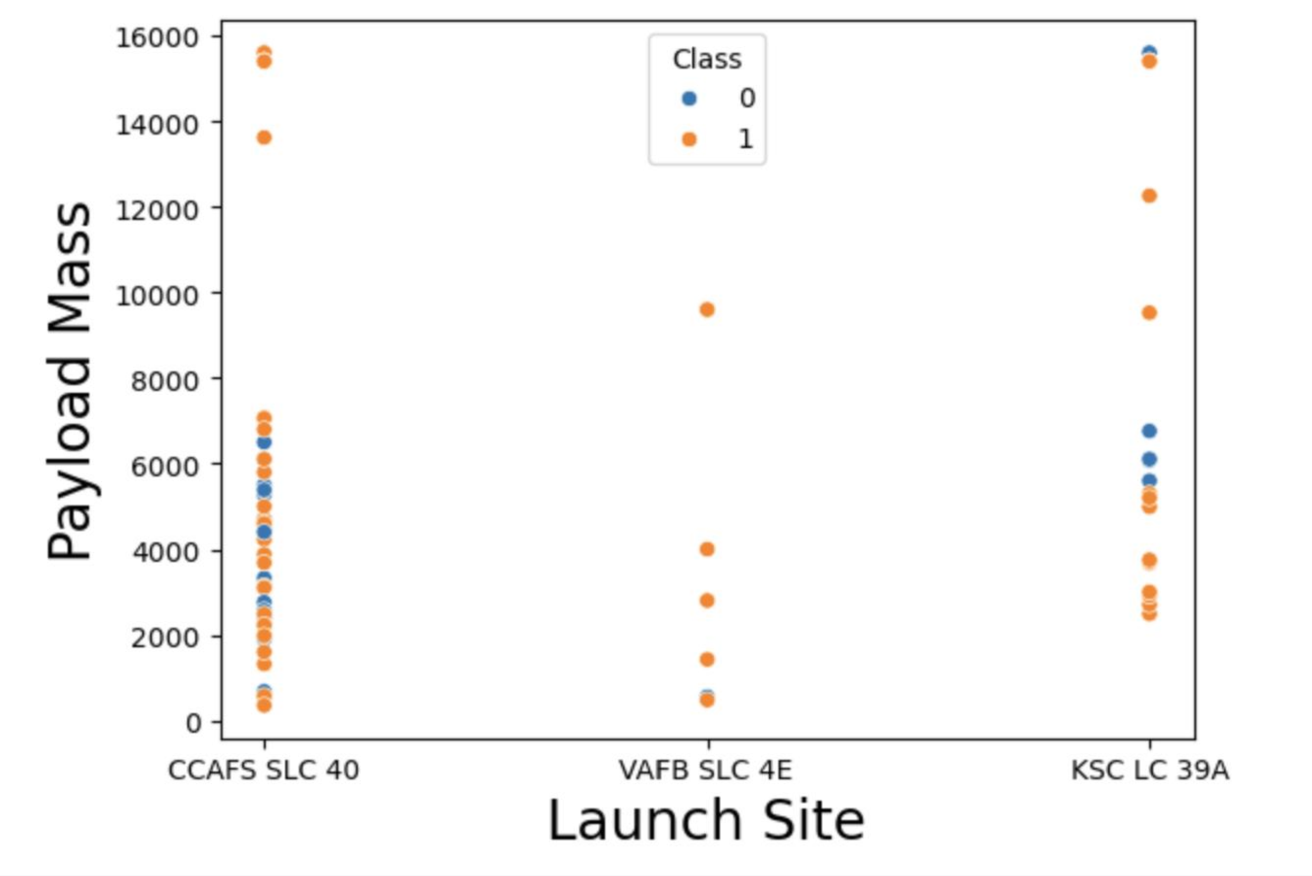
- For launch site CCAFS SLC 40 there does not seem to be a correlation between Flight Number and Launch Site.
- For VAFB SLC 4E and KSC LC 39A as the Flight Number increases the success rate increases. Moreover, there are no successful landings for flight numbers less than 22.



Payload vs. Launch Site

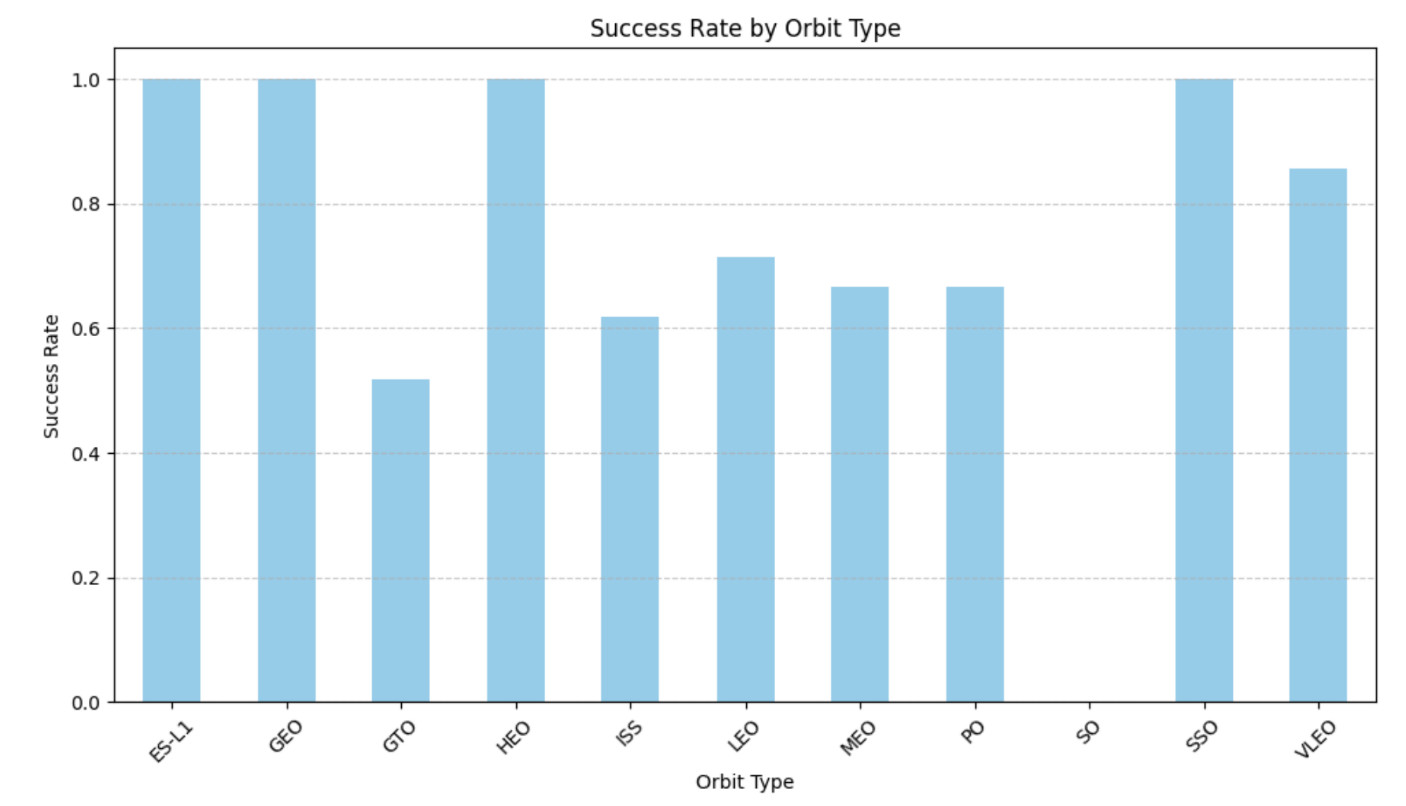
- For VAFB SLC 4E, there are no rockets launched for payload masses greater than 10000kg.

- There does not appear to be a direct correlation between the success of a landing based on Payload Mass increasing regardless of the Launch Site.



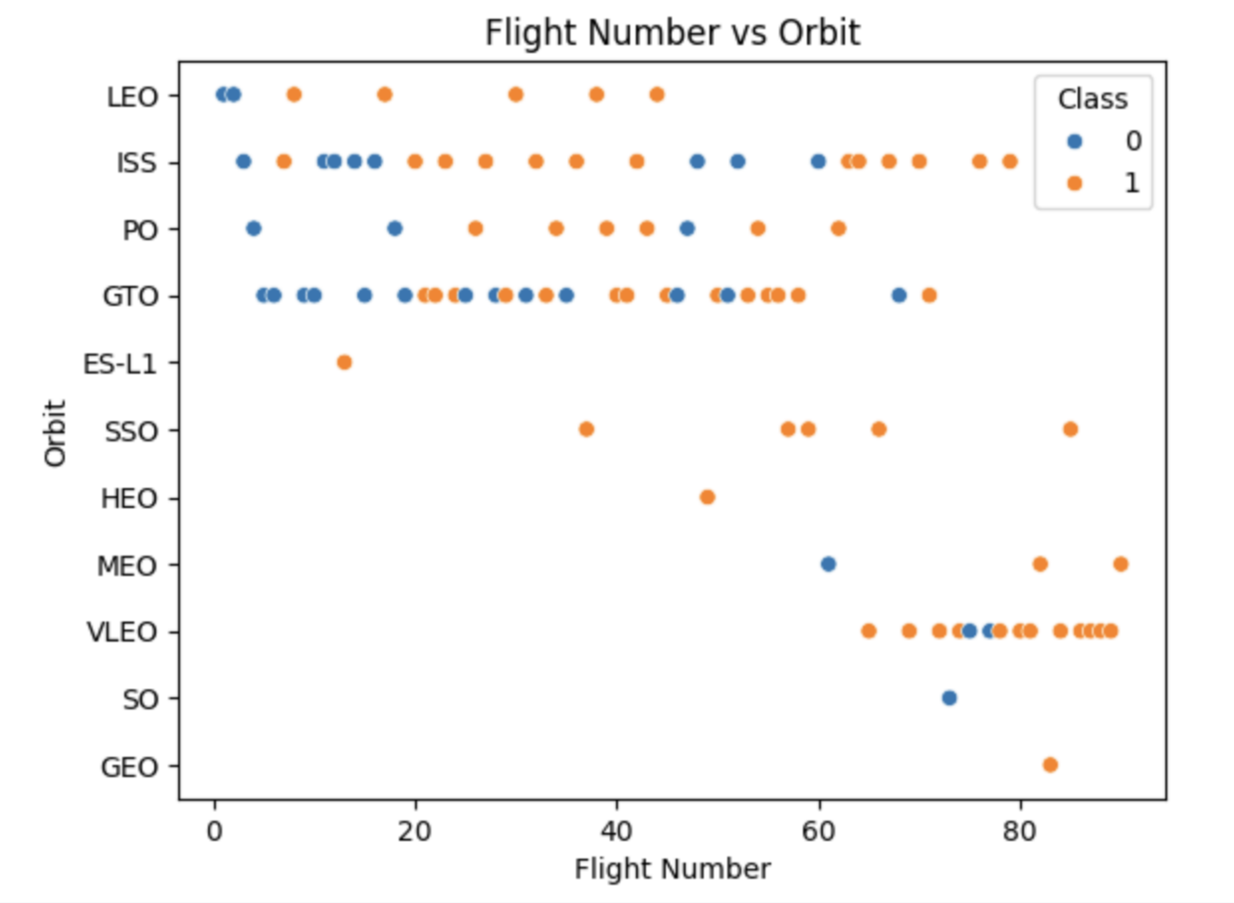
Success Rate vs. Orbit Type

- Orbits types ES-L1, GEO, HEO, and SSO had a perfect success rate.
- The lowest success rate was SO, which had a zero success rate.



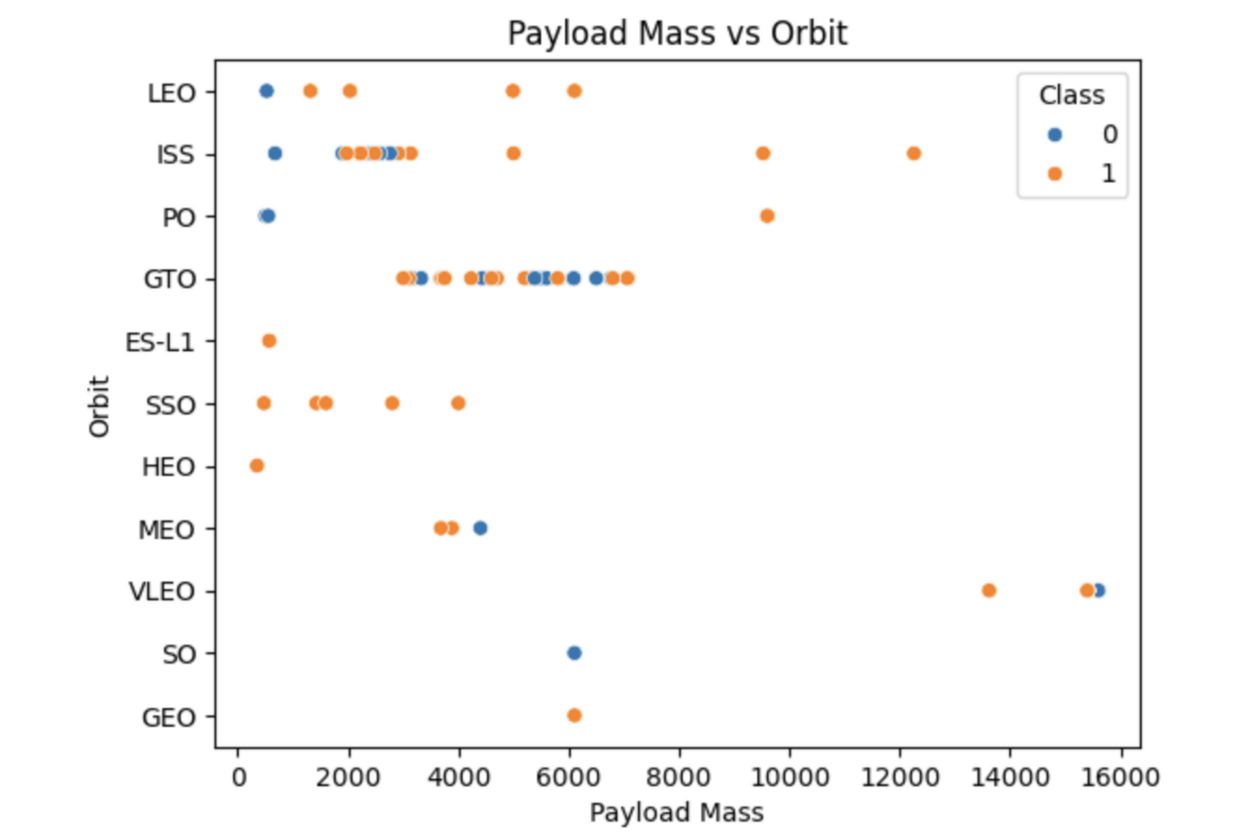
Flight Number vs. Orbit Type

- For the orbit type LEO, success rate is linked to Flight Number.
- For GTO there is no correlation between success rate and Flight Number.
- For ISS, the successful launches has flight numbers between 20-40, and 60+.



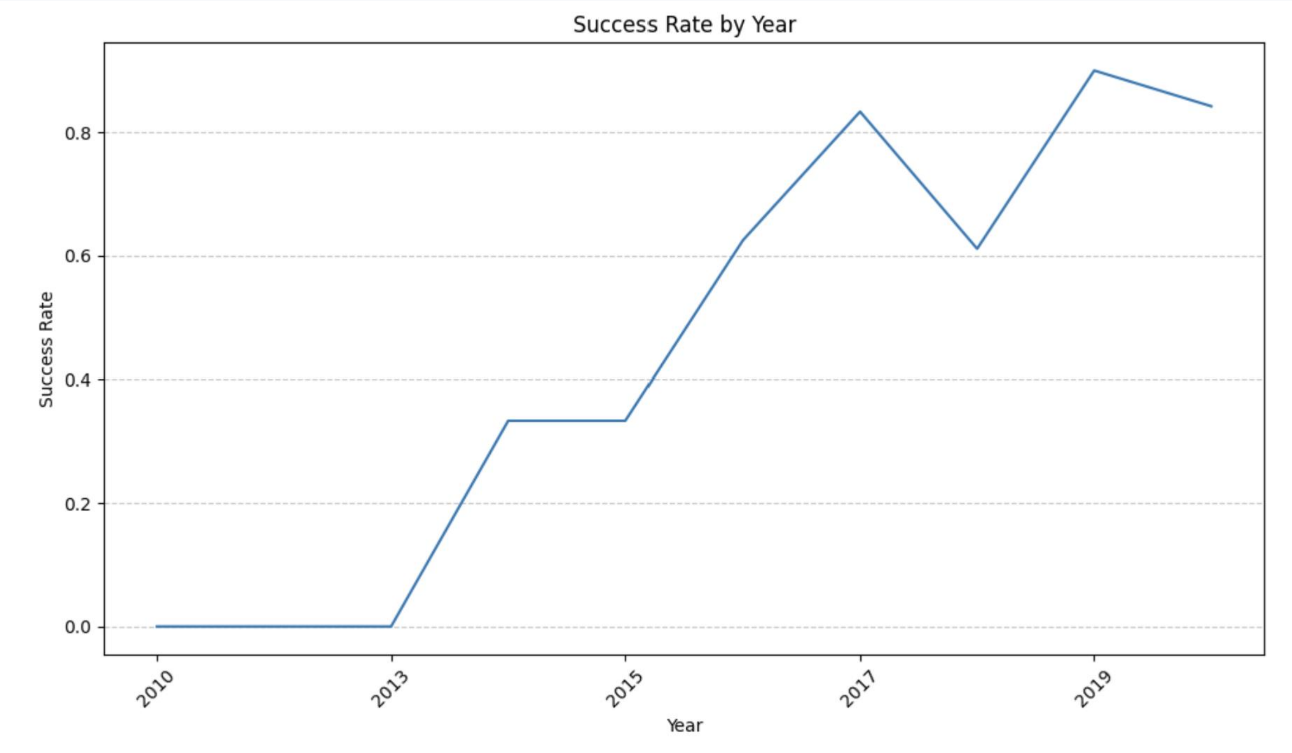
Payload vs. Orbit Type

- For PO, ISS, and LEO successful landings took place as Payload Mass increased.
- For GTO there does not seem to be a correlation between successful landings and increased Payload Mass.



Launch Success Yearly Trend

- The average success rate was constant between 2010 and 2013, as well as 2014 and 2015.
- The success rate increased from 2013 to 2017, and 2018 to 2019
- From 2017 to 2018 as well as 2019 onwards the success rate dropped.

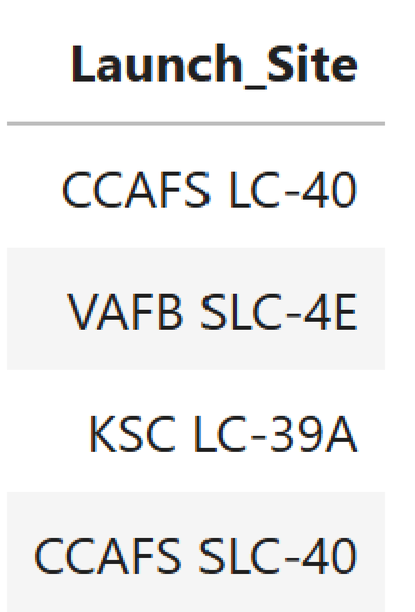


All Launch Site Names

- To find the names of the unique launch sites, the following query was used:

```
SELECT DISTINCT "Launch_Site"
FROM SPACEXTABLE;
```

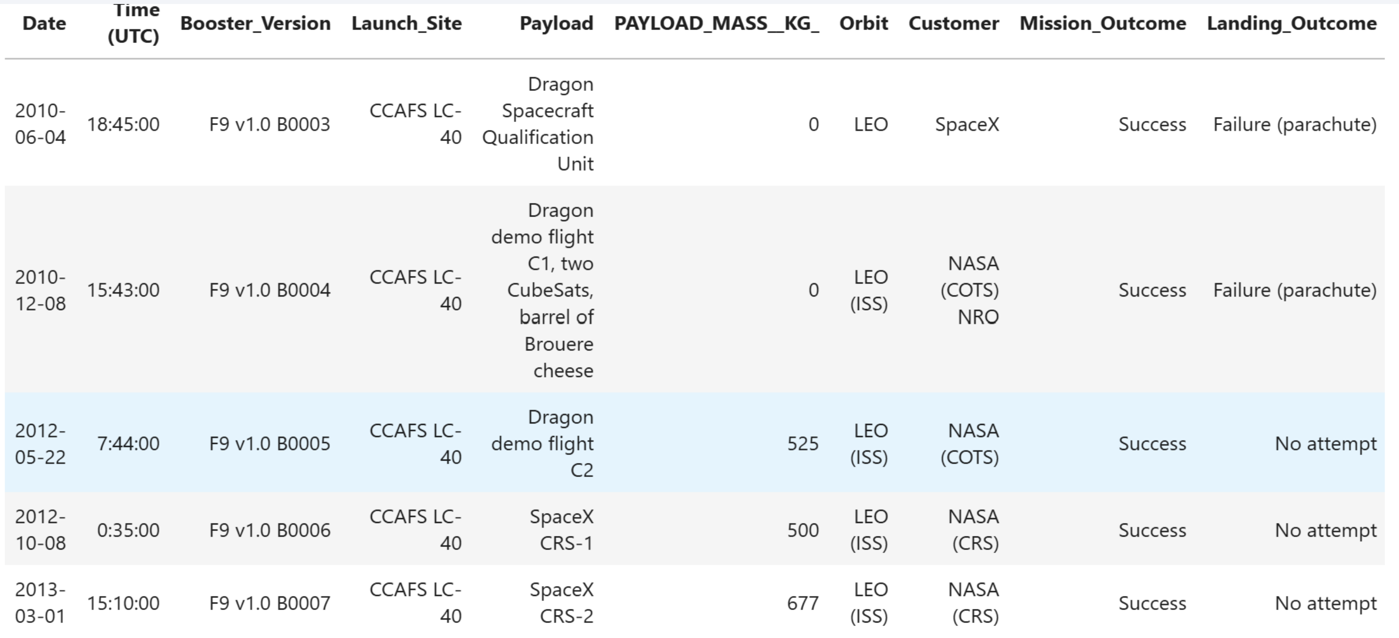
- These were the Launch Sites that were unique to the SpaceX launches.



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

Launch Site Names Begin with 'CCA'

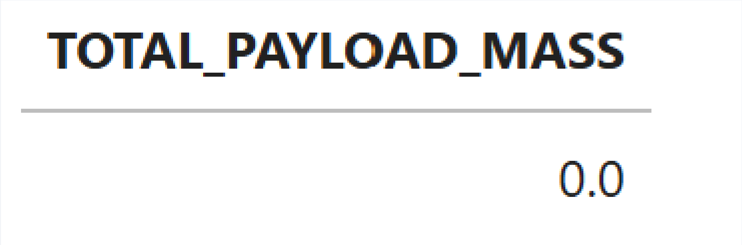
- To find 5 records where launch sites begin with `CCA`, the following query was used:
- **SELECT * FROM** SPACEXTABLE **WHERE** "Launch_Site" **LIKE** "CCA%" **LIMIT** 5;



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

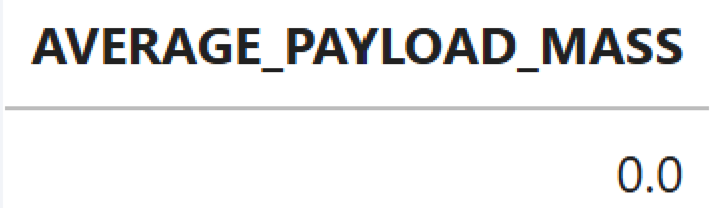
- To calculate the total payload carried by boosters from NASA, the following query was used
- `SELECT SUM("PAYLOAD_MASS_KG_") AS TOTAL_PAYLOAD_MASS`
- `FROM SPACEXTABLE`
- `WHERE "Customer" = 'NASA (CRS)';`
- The query result does not seem to be right here, even though the code executed was correct. There might be an issue with the PayLoad Mass column.



TOTAL_PAYLOAD_MASS
0.0

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
- **SELECT AVG**("PAYLOAD_MASS_KG_") **AS** AVERAGE_PAYLOAD_MASS **FROM** SPACEXTABLE **WHERE** "Booster_Version"='F9 v1.1';
- The query result does not seem to be right. There might be an issue with the PayLoad Mass column.

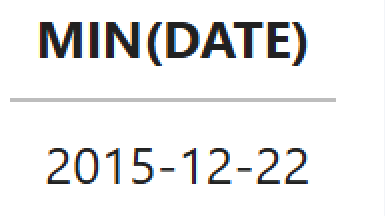


AVERAGE_PAYLOAD_MASS
0.0

First Successful Ground Landing Date

- To find the date of the first successful landing outcome on ground pad, the following query was used

-
  ```
  SELECT MIN(DATE) FROM SPACEXTABLE WHERE "Landing_Outcome"='Success (ground pad)';
  ```



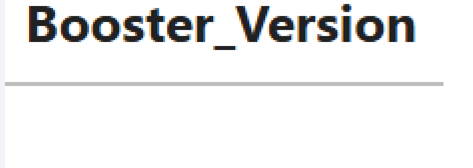
MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- To list 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

-
  ```
  SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE
  "Landing_Outcome"='Success (drone ship)' AND
  "PAYLOAD_MASS_KG_">4000 AND "PAYLOAD_MASS_KG_"<6000;
  ```

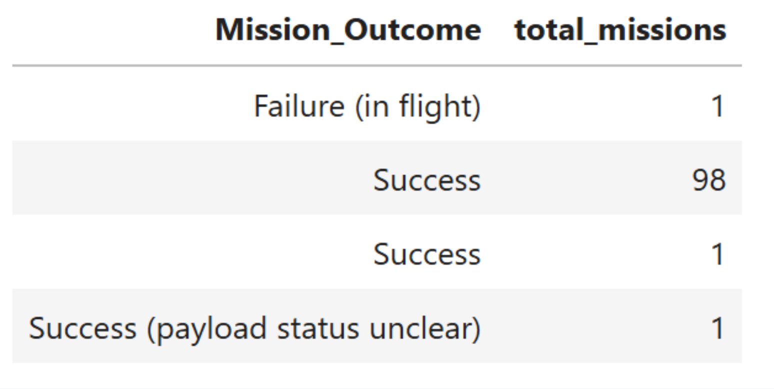
- The query result does not seem to be right. There might be an issue with the PayLoad Mass column.



Booster_Version

Total Number of Successful and Failure Mission Outcomes

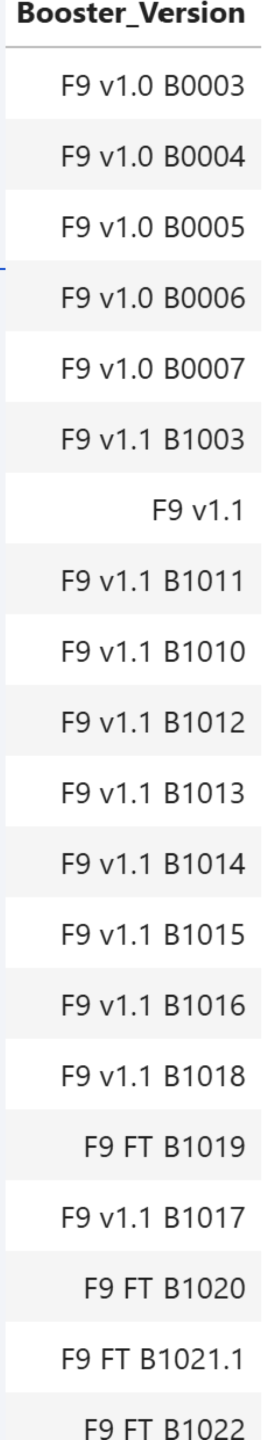
- To calculate the total number of successful and failure mission outcomes, the following query was used
- **SELECT** "Mission_Outcome", **COUNT**(*) **AS** total_missions
- **FROM** SPACEXTABLE
- **GROUP BY** "Mission_Outcome";
- There are a total of 100 successful missions, and one failed mission that occurred in flight.



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

Boosters Carried Maximum Payload

- To list the names of the booster which have carried the maximum payload mass, the following query was used
- SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS_KG_"=(SELECT MAX("PAYLOAD_MASS_KG_") FROM SPACEXTABLE);
- The list goes on, we have provided just a snippet.



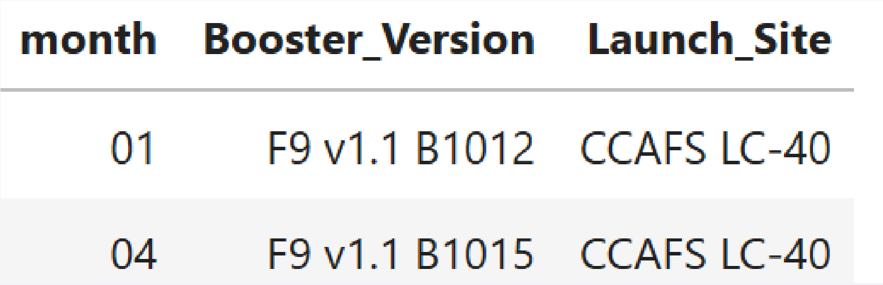
Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004
F9 v1.0 B0005
F9 v1.0 B0006
F9 v1.0 B0007
F9 v1.1 B1003
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1010
F9 v1.1 B1012
F9 v1.1 B1013
F9 v1.1 B1014
F9 v1.1 B1015
F9 v1.1 B1016
F9 v1.1 B1018
F9 FT B1019
F9 v1.1 B1017
F9 FT B1020
F9 FT B1021.1
F9 FT B1022

2015 Launch Records

- To list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, the following query was used

```
SELECT
    substr("Date", 6, 2) AS month,
    "Booster_Version","Launch_site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Failure (drone ship)'
  AND substr("Date", 0, 5) = '2015';
```

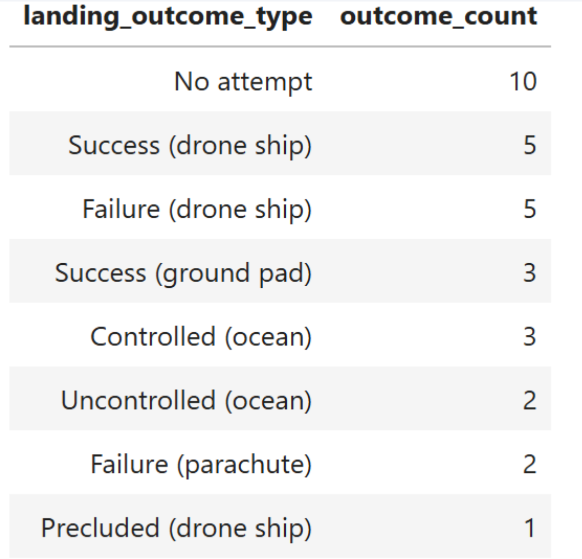
- There were two failed at occurred in January and April.



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

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

- To 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, the following query was used
- **SELECT**
-     "Landing_Outcome" **AS** landing_outcome_type,
-     **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**;
- The most popular landing outcome is no attempt which had ten outcome counts. The lowest ranking was precluded drone ship landings



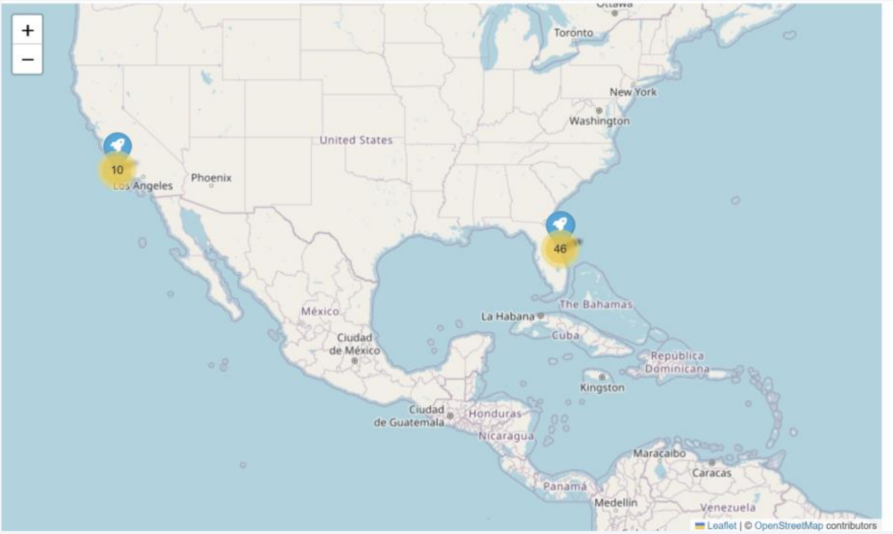
landing_outcome_type	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

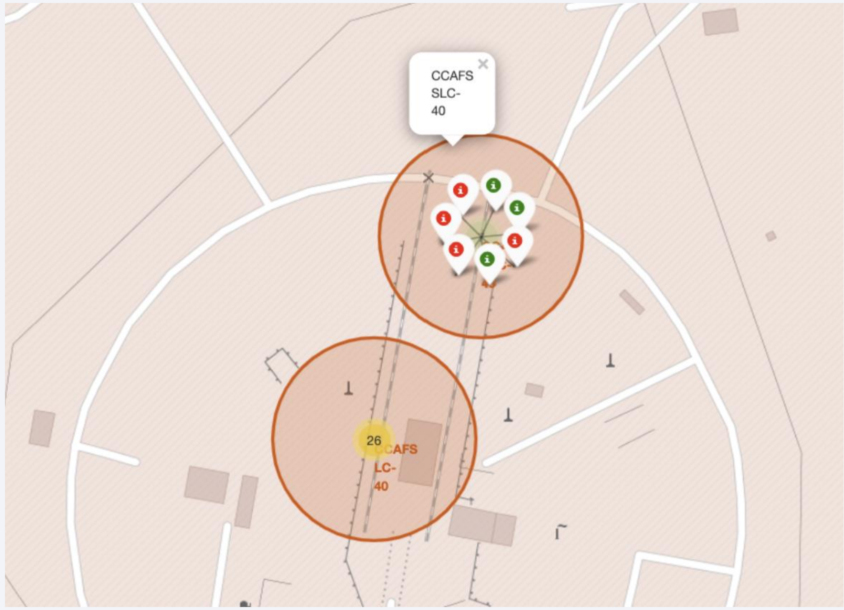
Launch Sites on a Global Map

- The launch sites are marked on the global folium map by blue space rockets. These launch sites represent the four main launch sites: CCASFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40.
- There are two clusters on the map in LA and FL, representing the number of launches in each launch sites. 10 launches in LA and 46 launches in FL in total.
- KSC LC-39A, CCAFS LC-40 and CCAFS SLC-40 are all in FL with 13, 26, and 7 launches respectively.
- VAFB SLC-4E is the only launch site in LA with 10 launches.



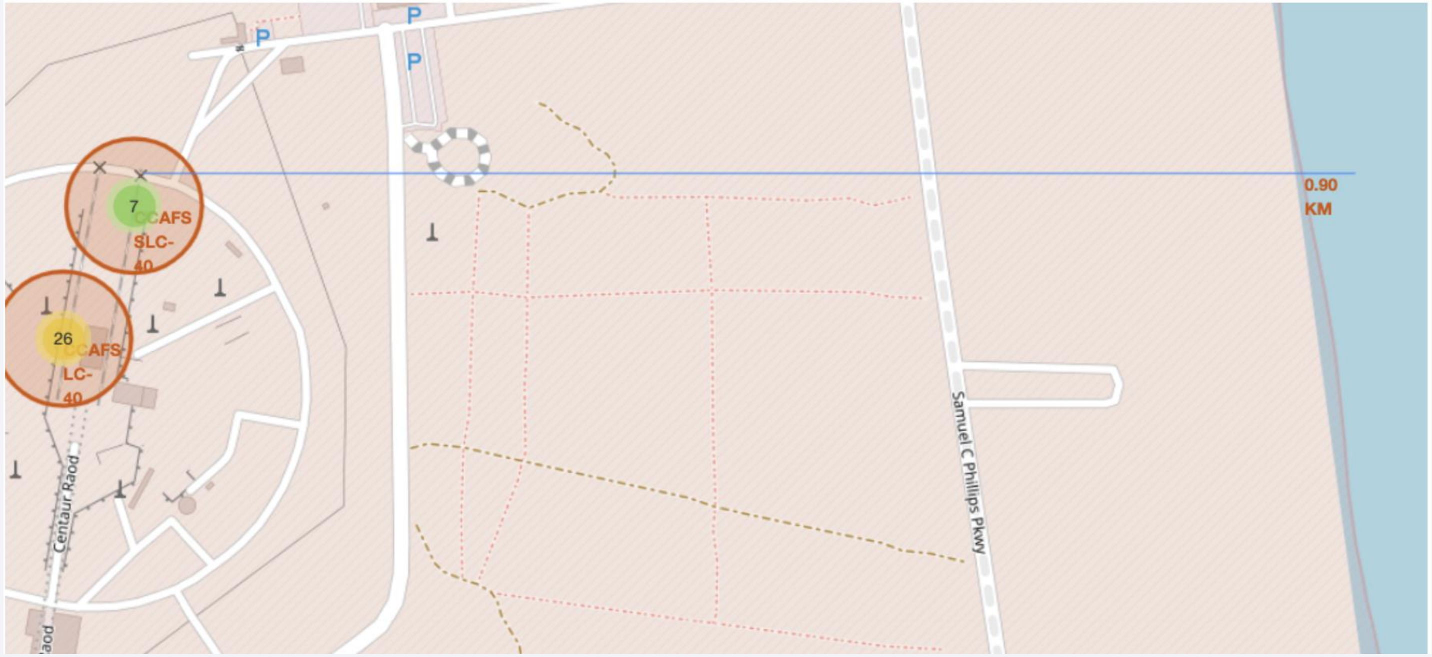
Successful Launch Outcomes on Folium Map

- The screenshot is an example of the color markers that were added to the launch clusters to indicate successful and unsuccessful launches.
- There were three successful landings in launch site CCAFS SLC-40 and four unsuccessful landings.



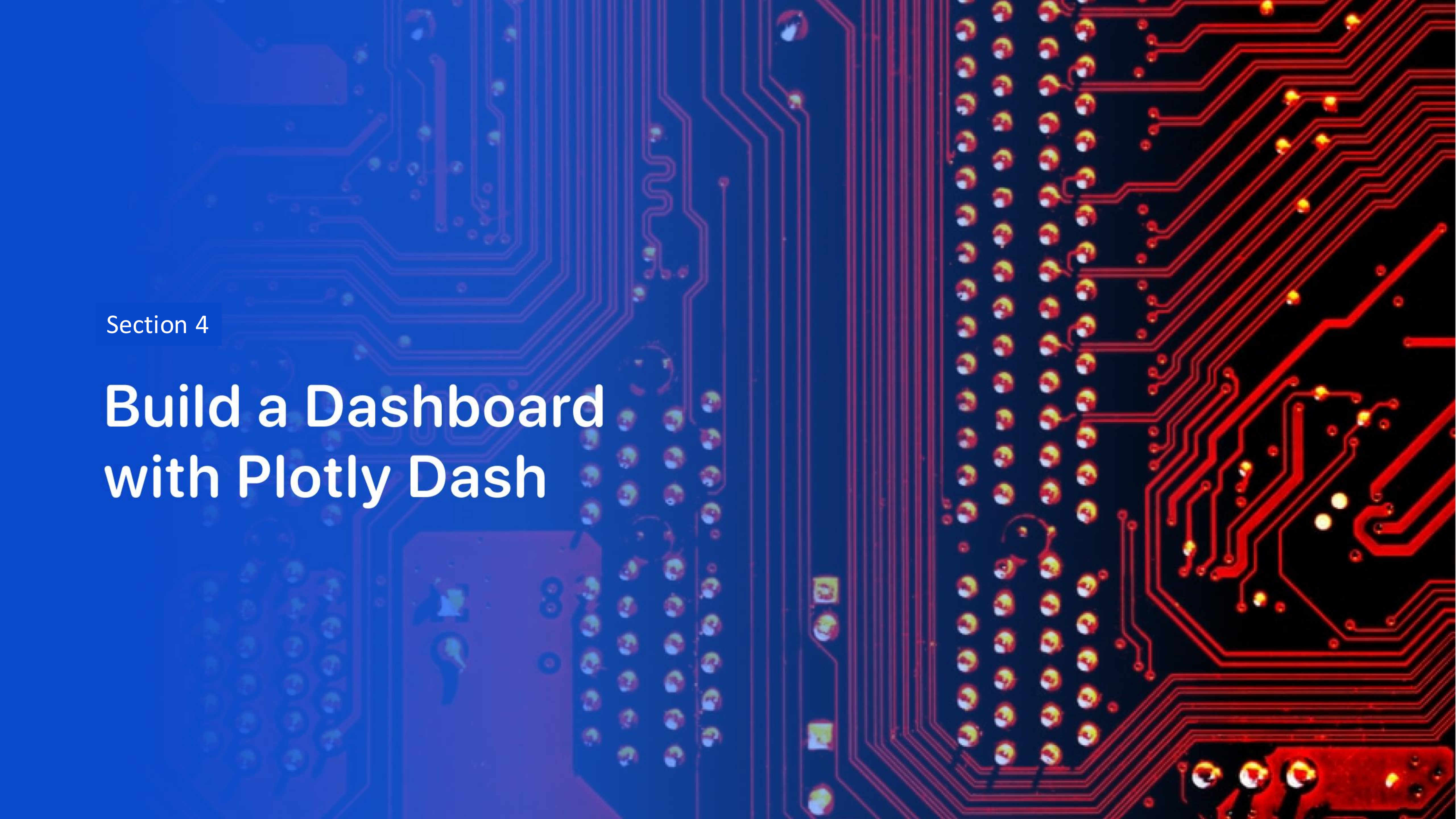
Distance to closest coastline on Folium Map

- The distance closest to the coast line was 0.90km from a launch recorded in the launch site CCAFS SLC-40.



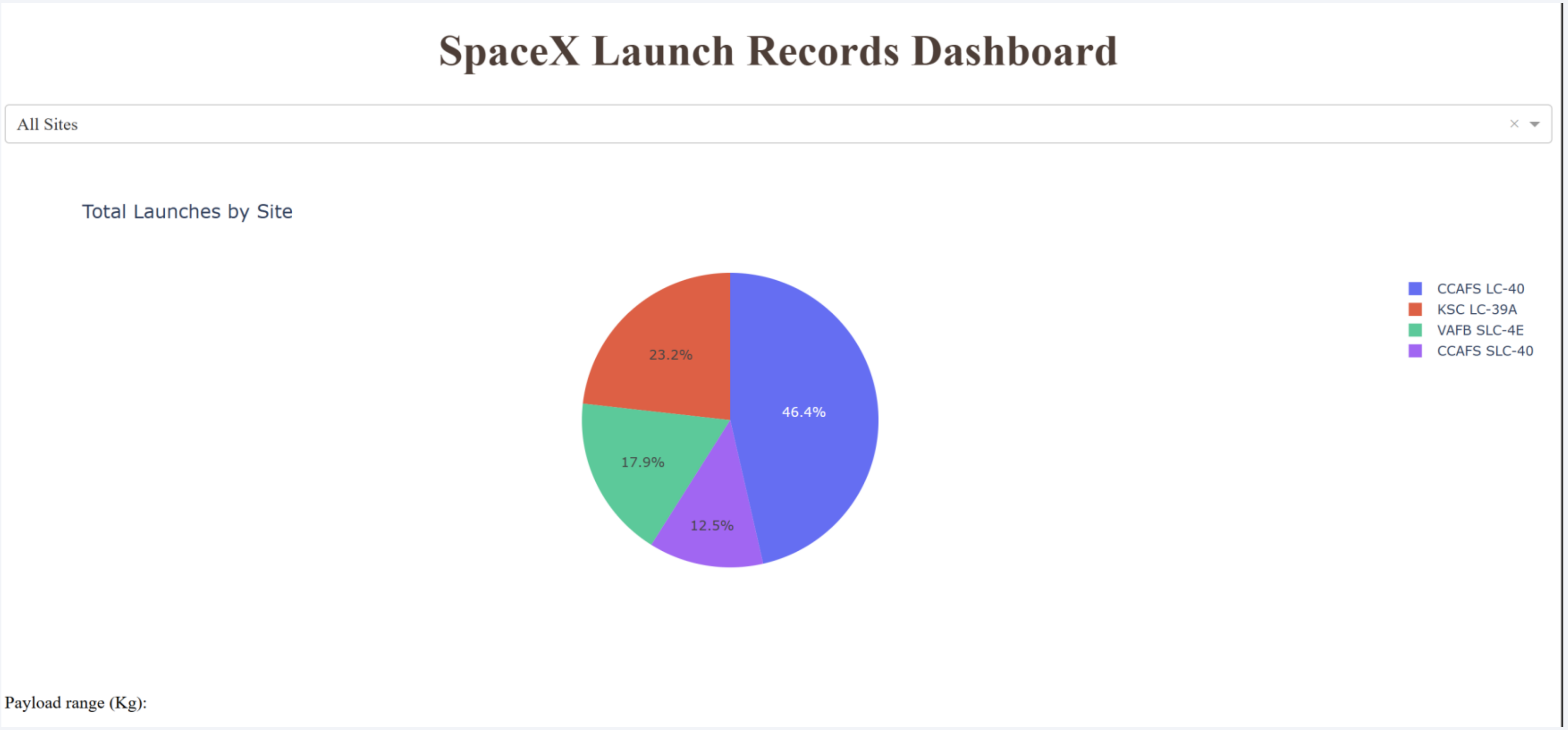
Section 4

Build a Dashboard with Plotly Dash



Launch Success Count for all Sites

- CCAFS LC-40 is the launch site with the most successful launches (46.4%), whilst CCAFS SLC-40 had the lowest count for successful launches (12.5%).



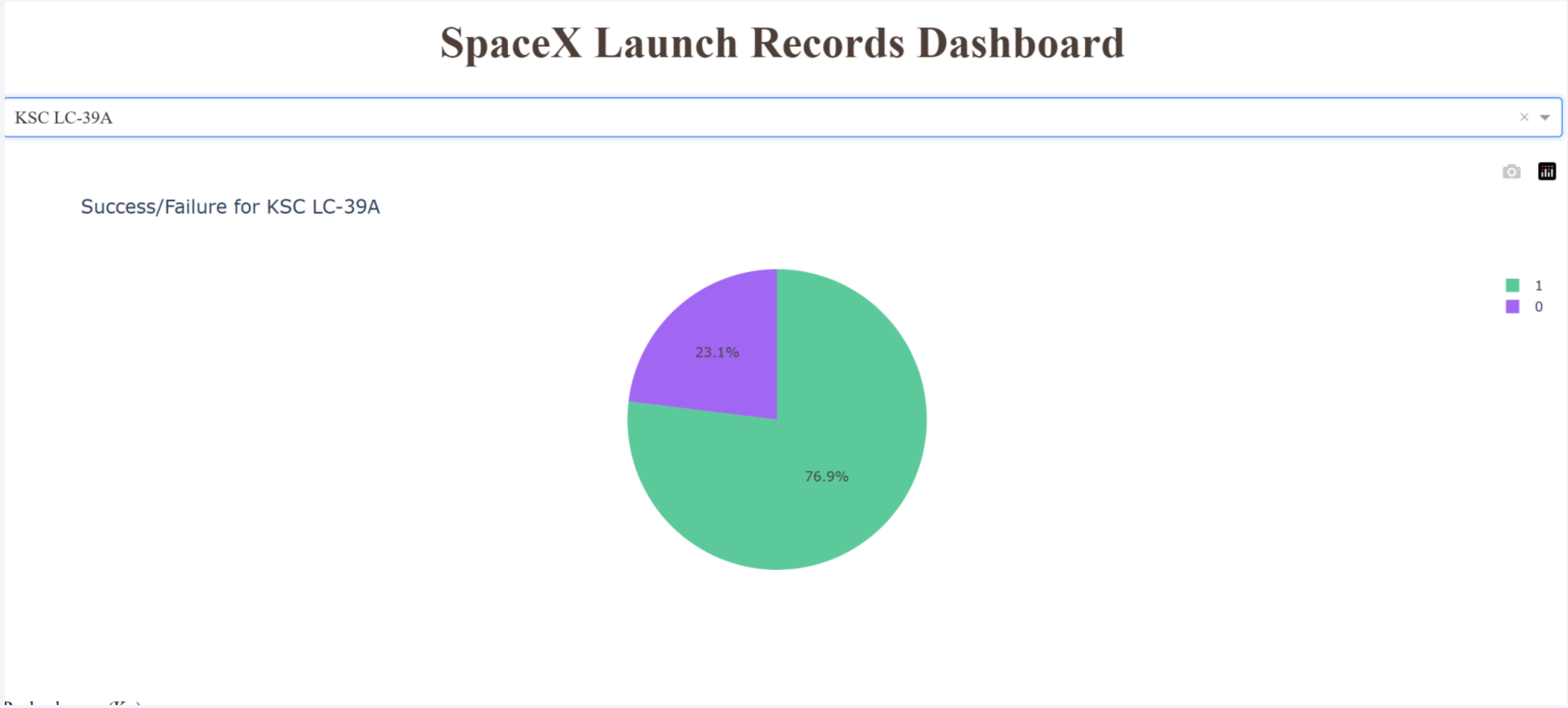
SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Launch Site with highest Launch Success Ratio

- KSC LC-39A is the launch site with the highest launch success ratio: 76.9% successful vs 23.1% failed launches out of all the other sites.

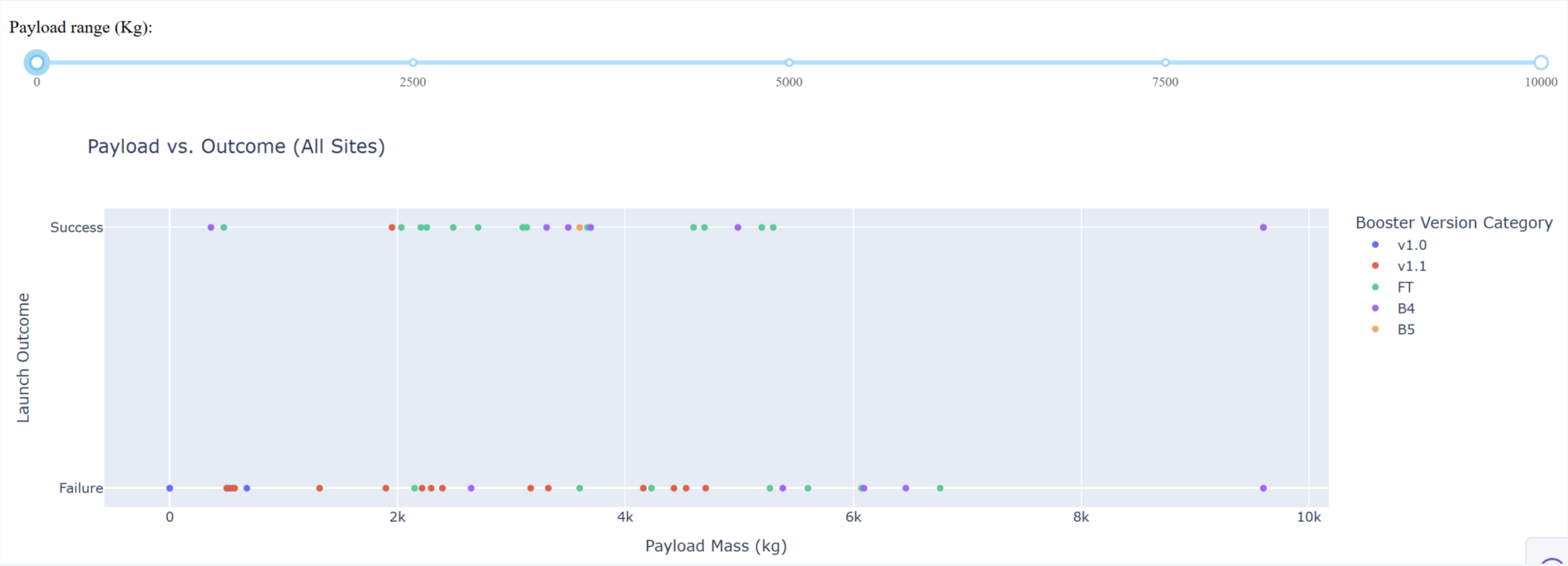


SpaceX Launch Records Dashboard

KSC LC-39A

Payload vs Launch Outcome Scatter Plots

- The FT booster has the highest success rate with payload mass in the range 2k-4k, whilst v1.1 has the lowest success rate for payload mass in the range 0-5k.

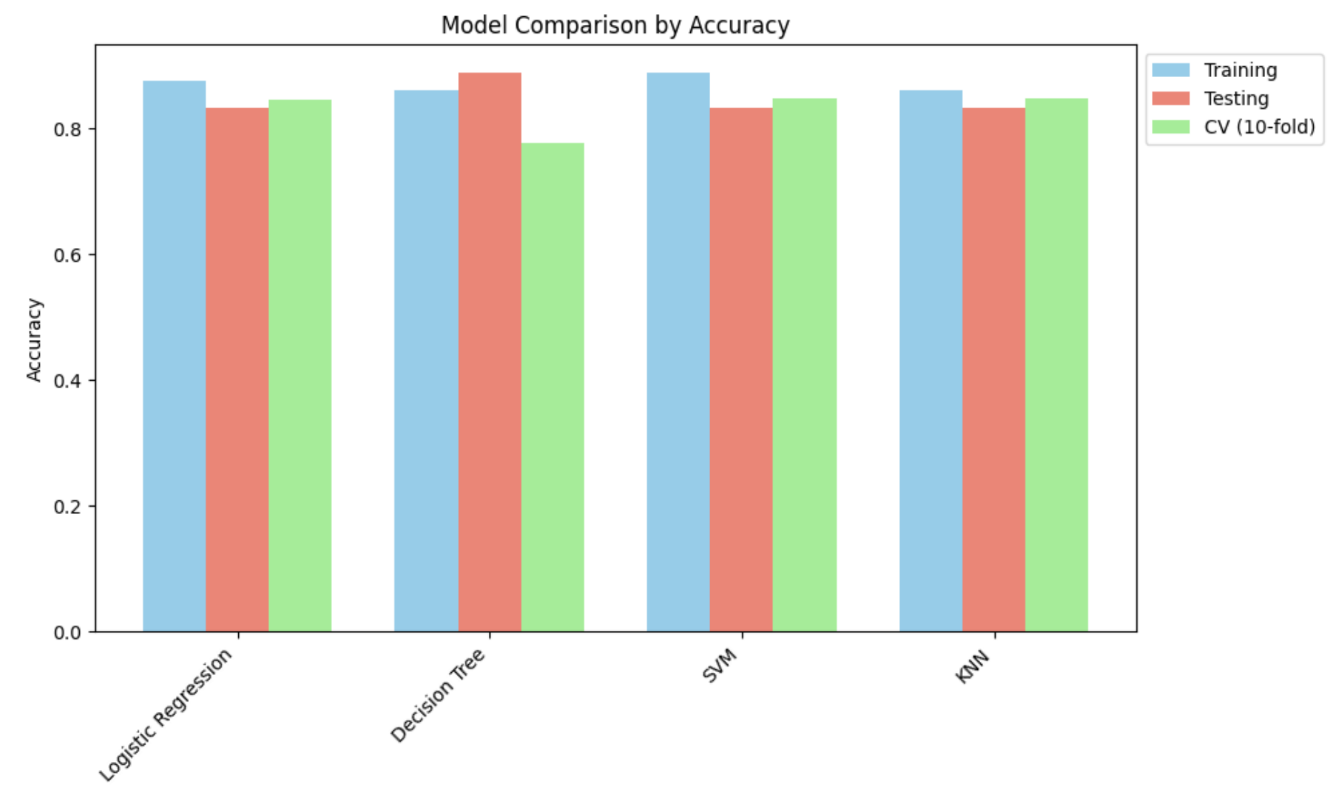


Section 5

Predictive Analysis (Classification)

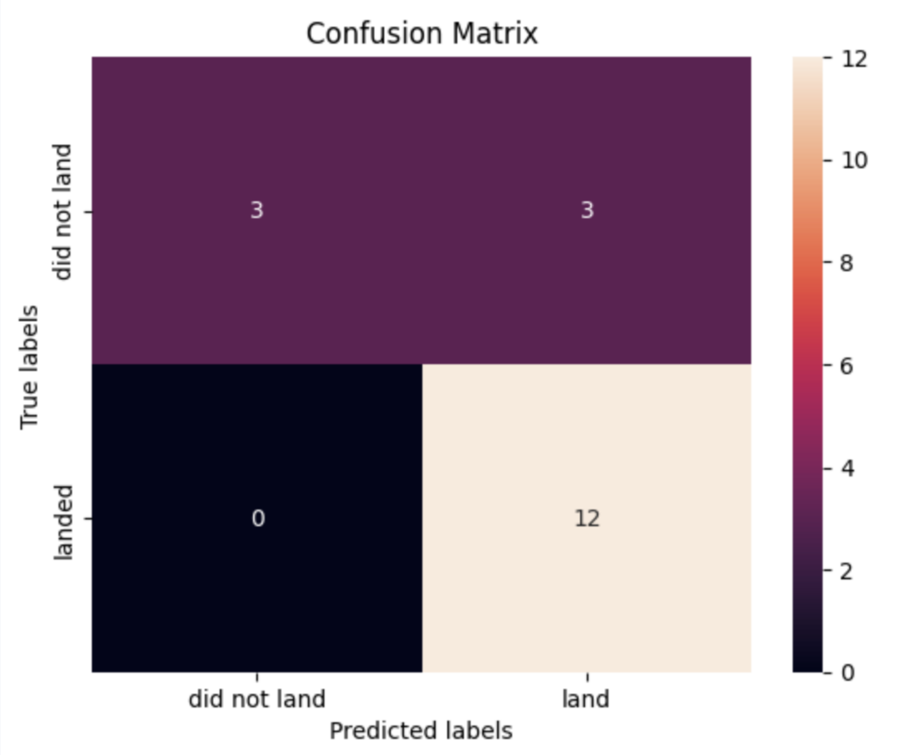
Classification Accuracy

- The best model is the SVM model since it has the highest cross validation accuracy of 0.848214 as well as the highest training accuracy of 0.88889, and a testing accuracy of 0.83333. This is the best model since all of these three accuracies are relatively close.



Confusion Matrix

- The confusion matrix of the SVM model shows that it can distinguish between the classes very well. There are 12 true positives, 3 false positives, and 3 true negatives, and zero false negatives. The only problem is the false negatives.



Conclusions

- The data shows SpaceX has successfully landed its Falcon9 rocket on numerous missions. This was evident in the data where the same booster version was recorded many times, meaning the booster was reused.
- Launches from 2018 onwards show higher and increasing landing success rate due to better landing technology and recovery process of the first stage.
- Lighter payloads seem to allow for the reusability of the first stage better than heavier payloads of more than 5,000kg.
- The launch site makes a difference: CCAFS LC-40 had the most successful missions whereas VAFB SLC-4E had the lowest. This could be due to different infrastructure which could affect the recovery process of the first stage.

Thank you!