Winning Space Race with Data Science

Keshav Mishra

Outline

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

Executive Summary

- **Summary of methodologies**
  Data Collection API & Web Scraping
  Data Wrangling
  EDA with SQL
  EDA with Data Visualization
  Building Interactive Map with Folium
  Building Dashboard with Plotly Dash
  Predictive Analysis (Classification)

- **Summary of all results**

  EDA results

  Interactive Analytics

  Predictive Analysis

Introduction

- **Project background and context**

  SpaceX has revolutionized space travel by reducing launch costs through the reuse of the Falcon 9 rocket's first stage. While traditional rocket launches can cost upwards of $165 million, SpaceX offers launches for around $62 million due to this reusability. The success of a mission—and the associated cost savings—greatly depends on whether the first stage lands successfully.

- **Problems we want to find answers**

  Can we predict whether the Falcon 9 first stage will land successfully, and how does this affect launch costs? By determining landing success, we can estimate launch expenses and provide valuable insights for competitors bidding against SpaceX.

Methodology

Methodology

Executive Summary

- Data collection methodology:

  SpaceX Rest API

  Web Scraping from Wikipedia

- Perform data wrangling

  We calculated the number of launches per site, occurrences of each orbit, mission outcomes per orbit type, and created a landing outcome label from the Outcome column.

  One Hot Encoding data fields for Machine Learning and data cleaning of null values and irrelevant columns

- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

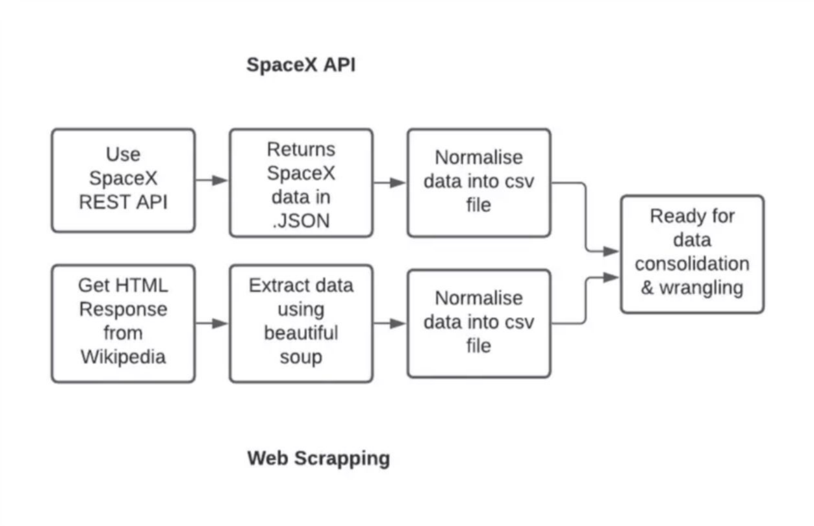
- Perform predictive analysis using classification models

  LR, KNN, SVM, DT models have been built and evaluated for the best classifier.

Data Collection

The following datasets was collected:

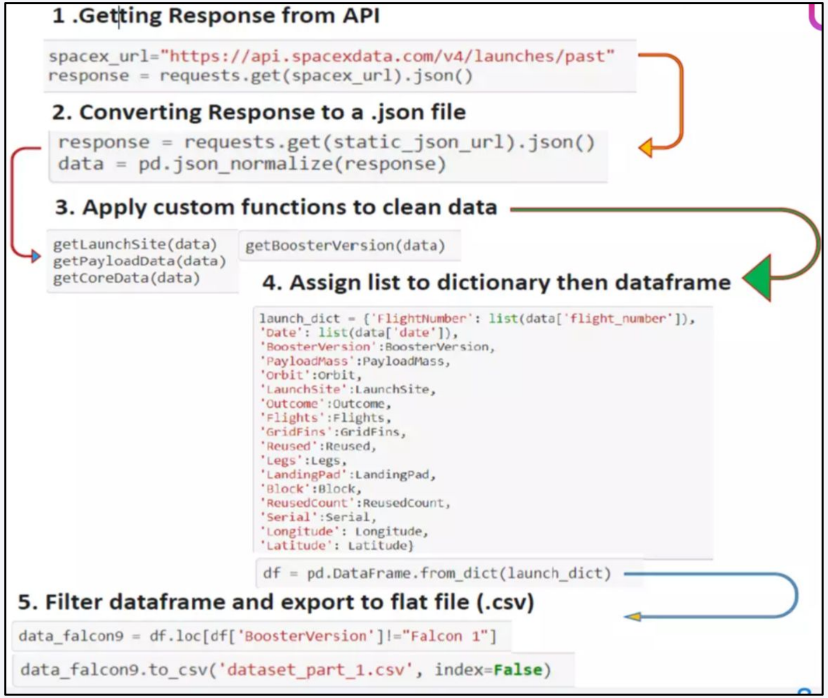
- SpaceX launch data that is gathered from the SpaceX REST API.
- This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and launch outcome.
- The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.
- Another popular data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup.



Data Collection - SpaceX API

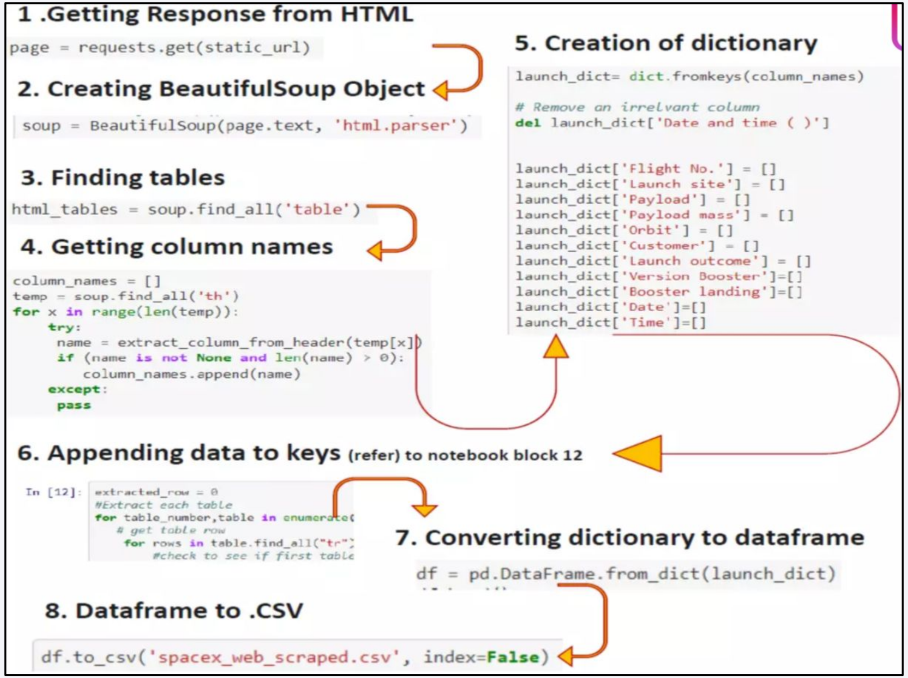
Data Collection with SpaceX REST calls.

- We make a get request to the SpaceX API. We also perform some basic data wrangling and formatting.
- It can be seen in detail in here :- https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex-data-collection-api.ipynb



Data Collection - Scraping

- We performed web scraping to collect historical Falcon 9 launch records from a Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches”.
- It can be seen in detail here :- https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex-webscraping.ipynb



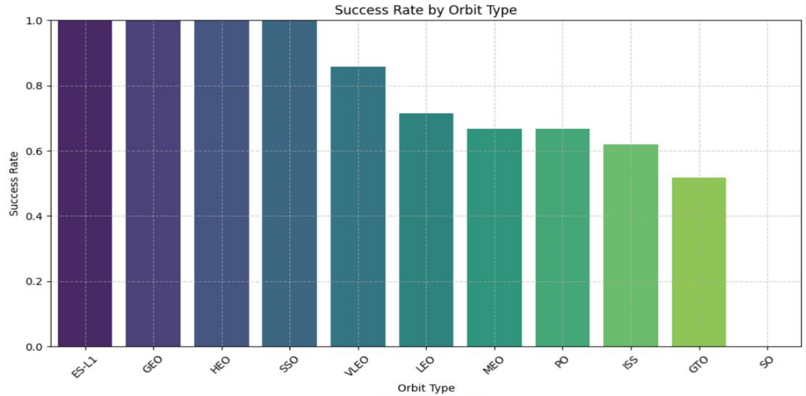
Data Wrangling

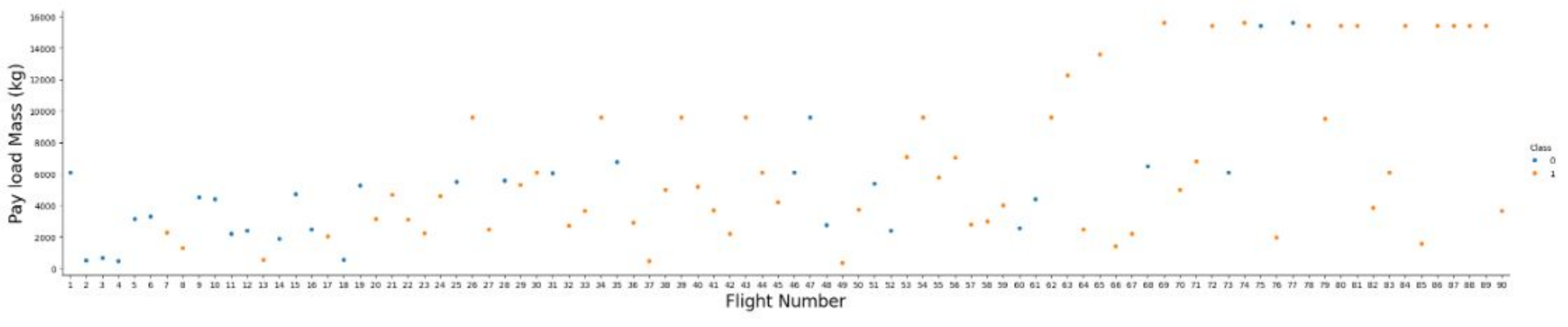
- Through a preliminary exploratory analysis identifying the transformations that are required in the data set to prepare them.
- We processed the landing data into valid tags for training the predictive models later.
- Training tags with "1" will mean the rocket landed successfully, and "0" means it was unsuccessful.
- It can be seen in detail here :- https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex-data%20wrangling.ipynb

EDA with Data Visualization

EDA to visualise relationship between:

- Flight Number and Launch Site.
- Payload and Launch Site.
- Success rate of each orbit type.
- Flight Number and Orbit type.
- Payload and Orbit type.
- Visualize the launch success yearly trend.





https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex-eda-viz.ipynb

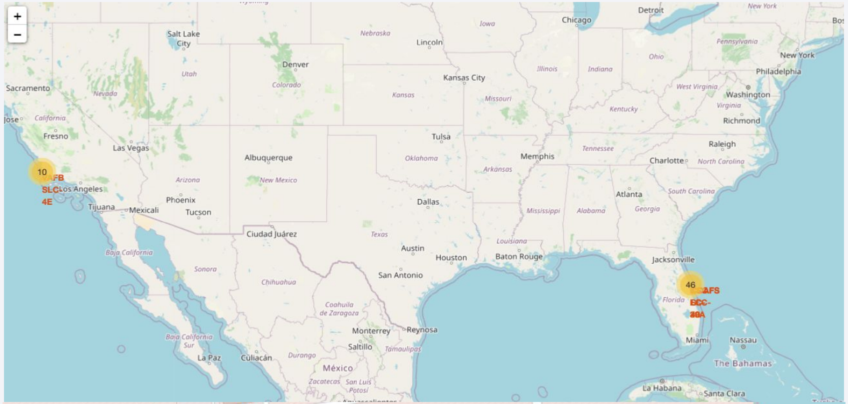
EDA with SQL

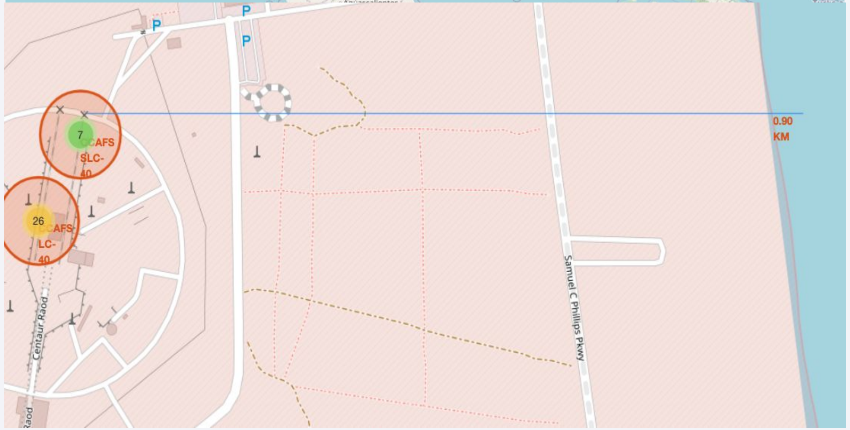
SQL queries performed include:

- Displaying the names of the unique launch sites in the space mission.
- Displaying 5 records where launch sites begin with the string "CCA".
- Displaying the total payload mass carried by boosters launched by NASA (CRS).
- Displaying average payload mass carried by booster version F9 v1.1.
- Listing the date where the successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster versions which have carried the maximum payload mass.
- List of Records of month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranking 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.

https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex-eda-sql.ipynb

Build an Interactive Map with Folium





- Markers, circles, lines and marker clusters were used with Folium Maps.
- Indications of each element:
- Markers indicate points like launch sites.
- Circles indicate highlighted areas around specific coordinates,like NASA Johnson Space Center.
- Marker clusters indicates groups of events in each coordinate,like launches in a launch site.
- Lines are used to indicate distances between two coordinates.
- Map markers have been added to the map with aim to finding an optimal location for building a launch site
https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Elements:

- Dropdown list for the launch site.
- RangeSlider for selecting the payload mass.
- PieChart: for showing the success rate of each launch site, or showing the number of successful landing outcomes.
- Scatterplot: Show success/failure by payload and booster version.

Findings:

- Which site has the largest successful launches? KSC LC-39A.
- Which site has the highest launch success rate? KSC LC-39A (success rate76.9%).
- Which payload range(s) has the highest launch success rate? 2000-4000.
- Which payload range(s) has the lowest launch success rate? 6000-8000.
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate ?

https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We create a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.
- Perform exploratory Data Analysis and determine Training Labels
- Create a column for the class
- Standardize the data
- Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data.

https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex_machine_learning_prediction.ipynb

Results

Exploratory Data Analysis Result

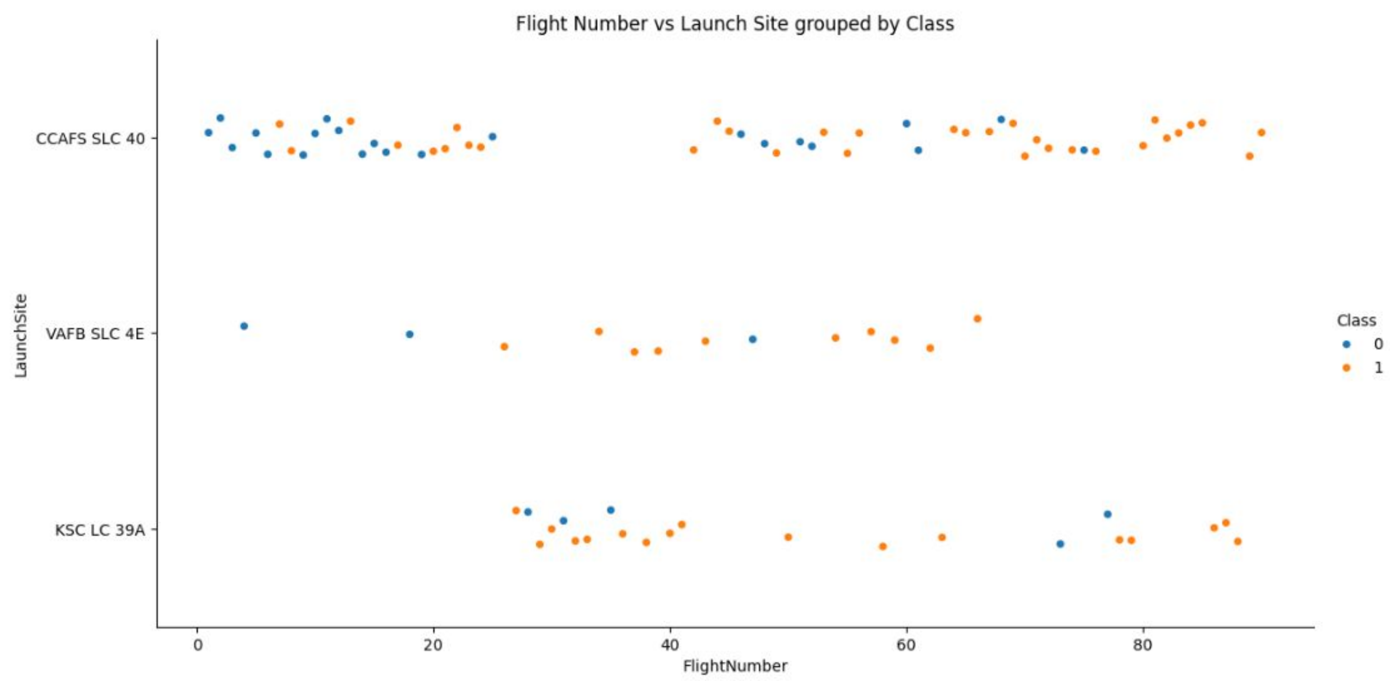
Interactive Analytics Demo in Screenshot

Predictive Analysis Result

Section 2

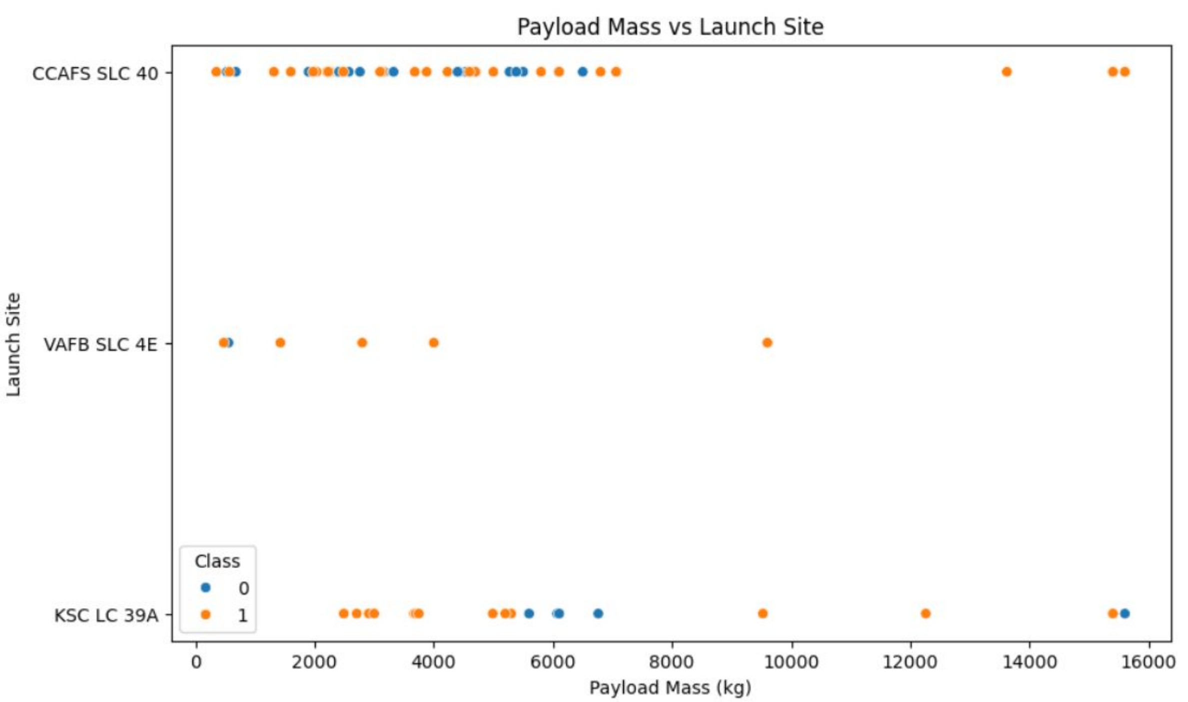
Insights drawn from EDA

Flight Number vs. Launch Site



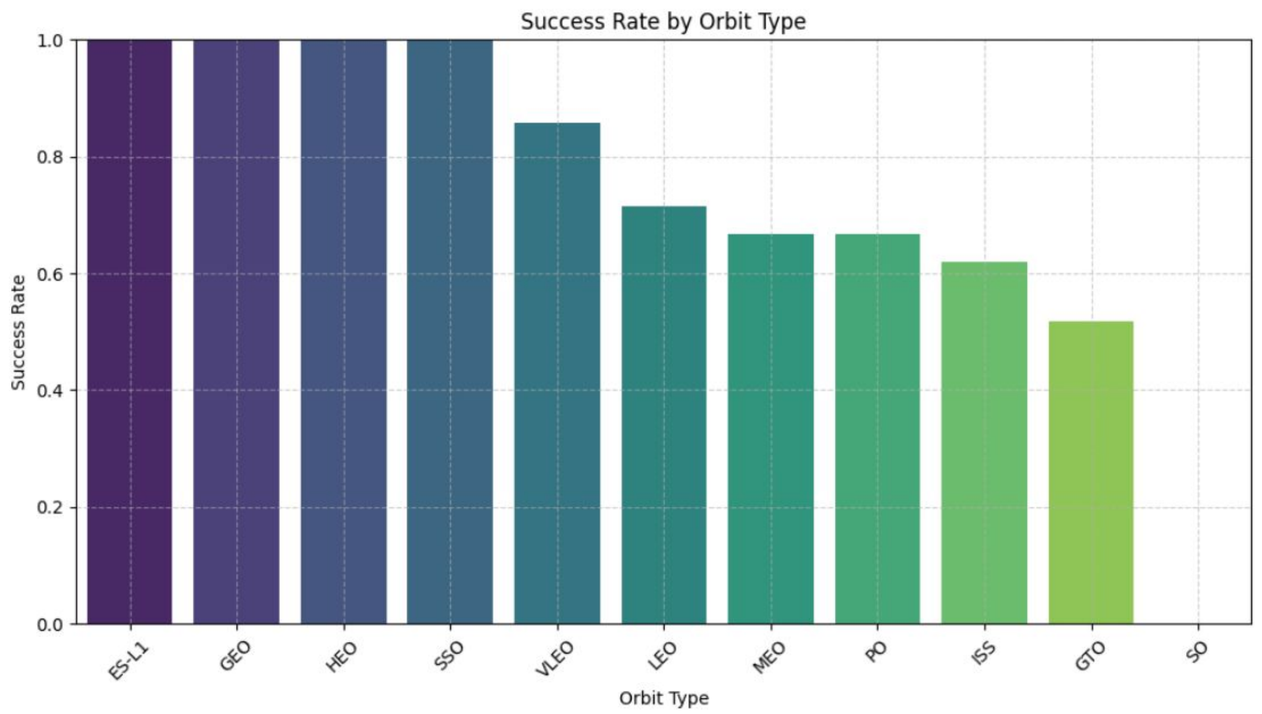
As flight numbers increase, the success rate (Class 1) improves, especially at CCAFS SLC 40, indicating growing reliability over time. KSC LC 39A shows high success even with fewer launches, while VAFB SLC 4E has limited but mostly successful launches.

Payload vs. Launch Site



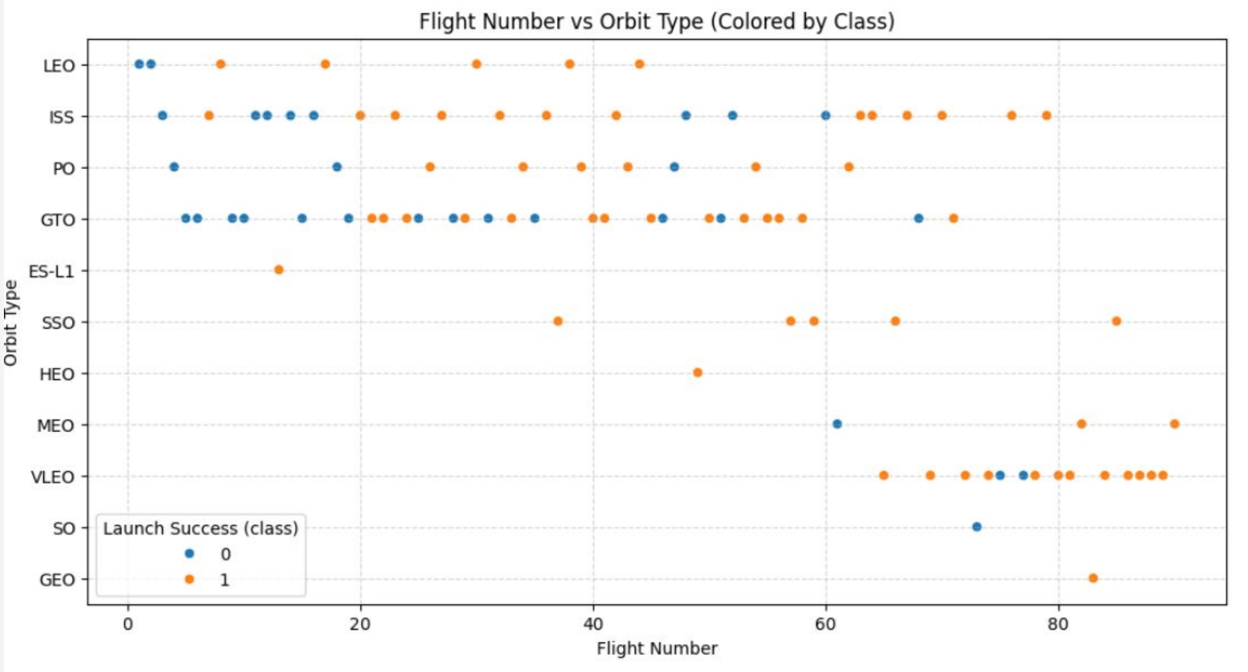
For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).
Payload mass alone may not strongly influence landing success

Success Rate vs. Orbit Type



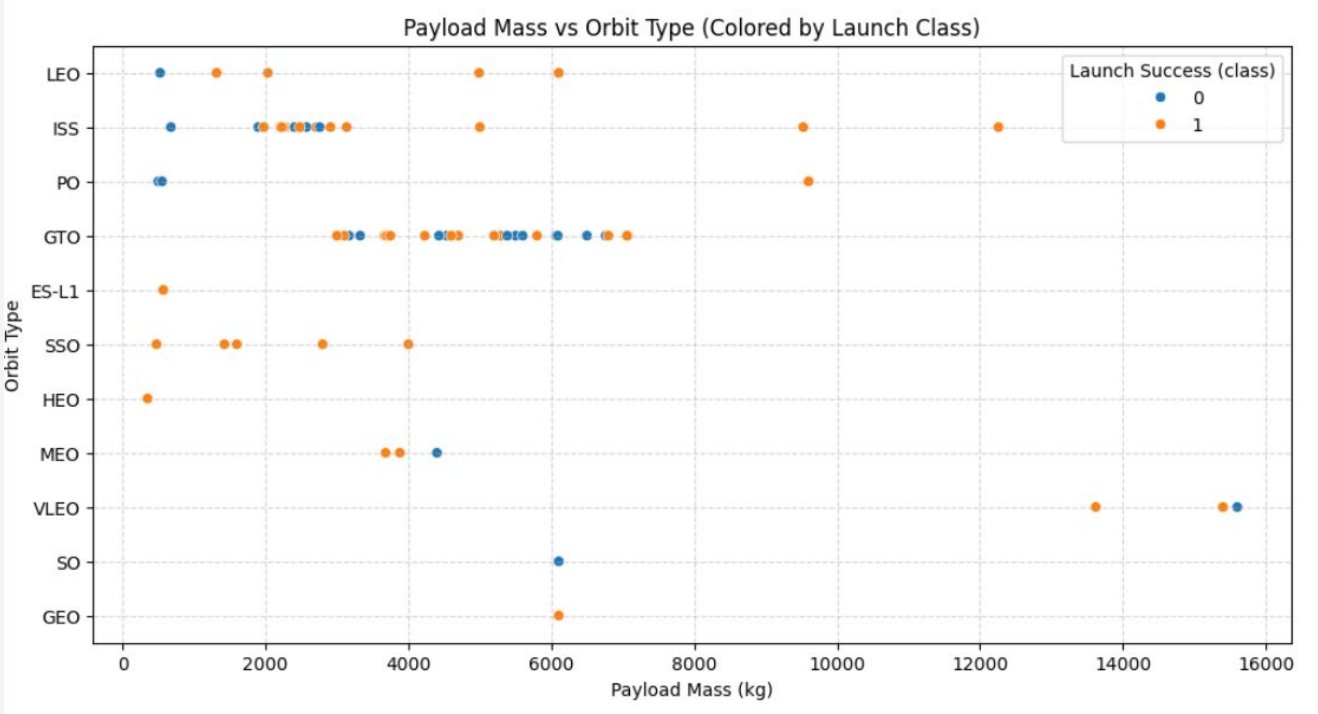
Landing success rates soar for higher orbits like GEO, HEO and ES-L1 and lower orbits such as SSO and graph reveal greater variability—highlighting specific altitude and mission complexity critically shape Falcon 9's landing reliability.

Flight Number vs. Orbit Type



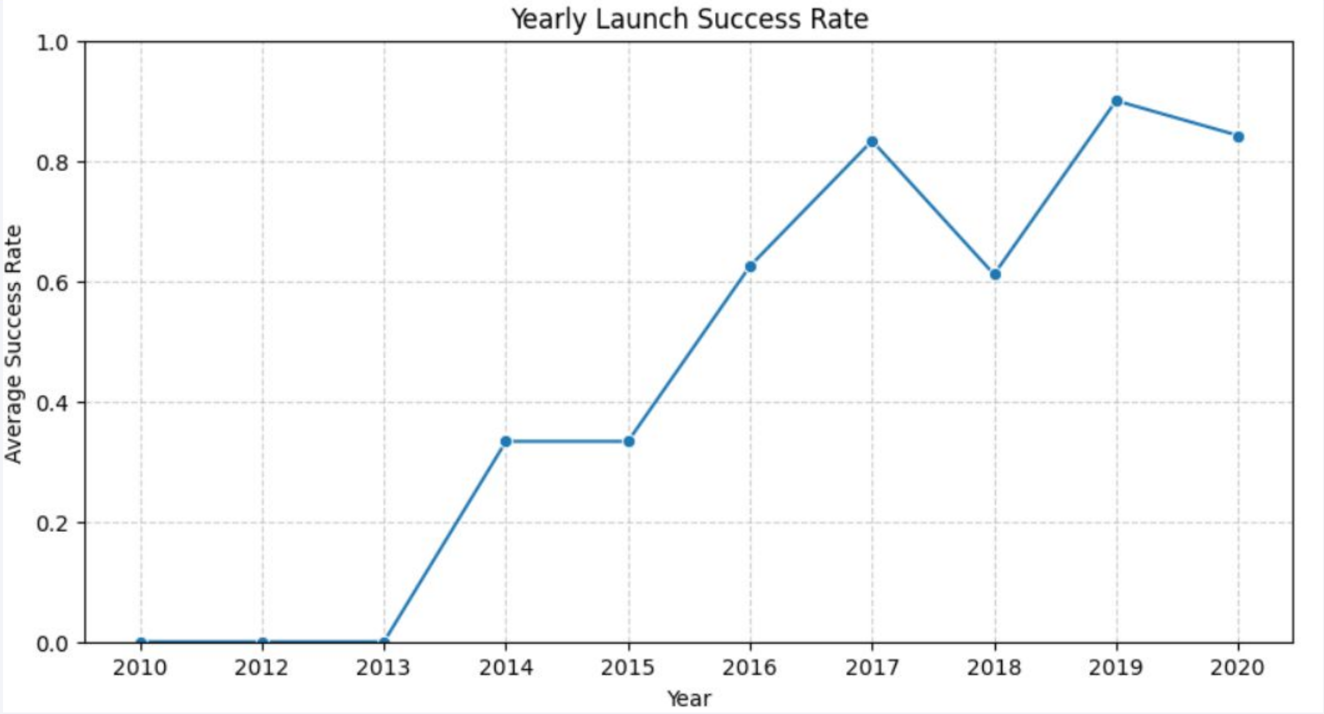
In LEO missions, landing success improves with flight experience, while in GTO missions, success appears unaffected by flight number—highlighting the role of orbit complexity over operational maturity. Also trend is observed of shifting to VLEO over period.

Payload vs. Orbit Type



Successful landings with heavier payloads are more frequent in LEO, Polar, and ISS missions, while GTO missions show no clear pattern.

Launch Success Yearly Trend



SpaceX's launch success rate has shown a strong upward trend since 2013, peaking in 2019, reflecting rapid technological improvements and operational maturity over time.

All Launch Site Names

SQL Query

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

Result:-

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

- CCAFS SLC 40 – Cape Canaveral Space Launch Complex 40
- KSC LC 39A – Kennedy Space Center Launch Complex 39A
- VAFB SLC 4E – Vandenberg Space Launch Complex 4E

Launch Site Names Begin with 'CCA'

SQL Query

```
%sql SELECT * FROM SPACEXTBL 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

SQL Query

```
%%sql SELECT SUM(Payload_Mass__kg_) AS Total_Payload_Mass
    FROM SPACEXTBL
    WHERE Customer = 'NASA (CRS)';
```

Avg_Payload
2928.4

Average Payload Mass by F9 v1.1

SQL Query

```
%%sql SELECT AVG(Payload_Mass__kg_) AS Avg_Payload
      FROM SPACEXTBL
      WHERE Booster_Version = 'F9 v1.1';
```

Avg_Payload
2928.4

First Successful Ground Landing Date

SQL Query

```
%%sql SELECT MIN(Date) AS First_Successful_Ground_Landing
        FROM SPACEXTBL
        WHERE Landing_Outcome = 'Success (ground pad)';
```

First_Successful_Ground_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query

```
%%sql SELECT Booster_Version
        FROM SPACEXTBL
        WHERE Landing_Outcome = 'Success (drone ship)'
           AND Payload_Mass__kg_ > 4000
            AND Payload_Mass__kg_ < 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

SQL Query

```
%%sql SELECT Mission_Outcome, COUNT(*) AS Total
    FROM SPACEXTBL
    GROUP BY Mission_Outcome;
```

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

Boosters Carried Maximum Payload

SQL Query

```
%%sql SELECT Booster_Version, Payload_Mass__kg_
        FROM SPACEXTBL
        WHERE Payload_Mass__kg_ = (
                SELECT MAX(Payload_Mass__kg_)
                FROM SPACEXTBL
         );
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

SQL Query

```
%%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_Name,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTBL
WHERE substr(Date, 1, 4) = '2015'
  AND Landing_Outcome = 'Failure (drone ship)';
```

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

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

SQL Query

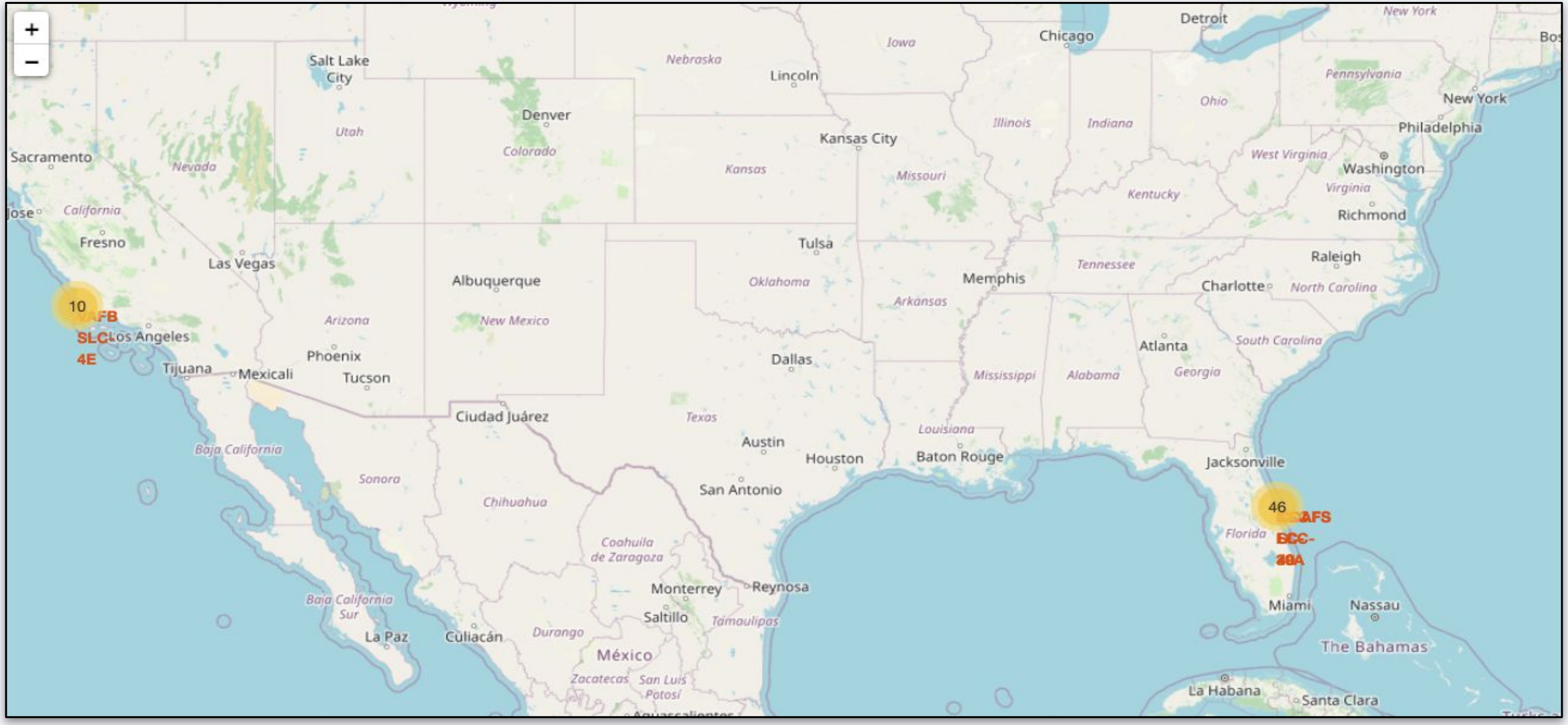
```
%%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count
    FROM SPACEXTBL
    WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY Landing_Outcome
    ORDER BY Outcome_Count DESC;
```

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 4

Launch Sites Proximities Analysis

All Launch Sites Marked on Map



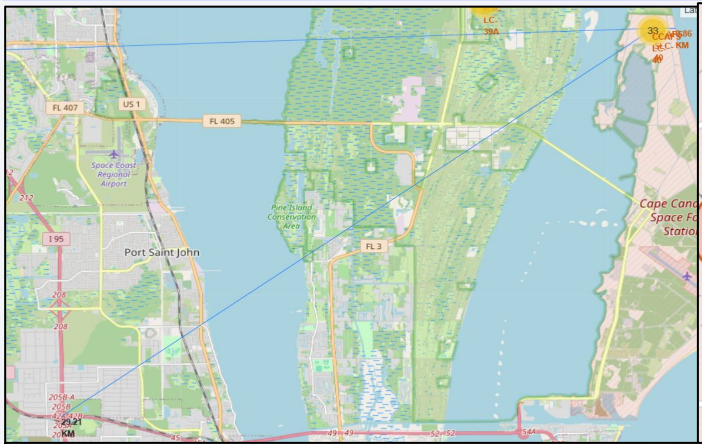
https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex_launch_site_location.ipynb

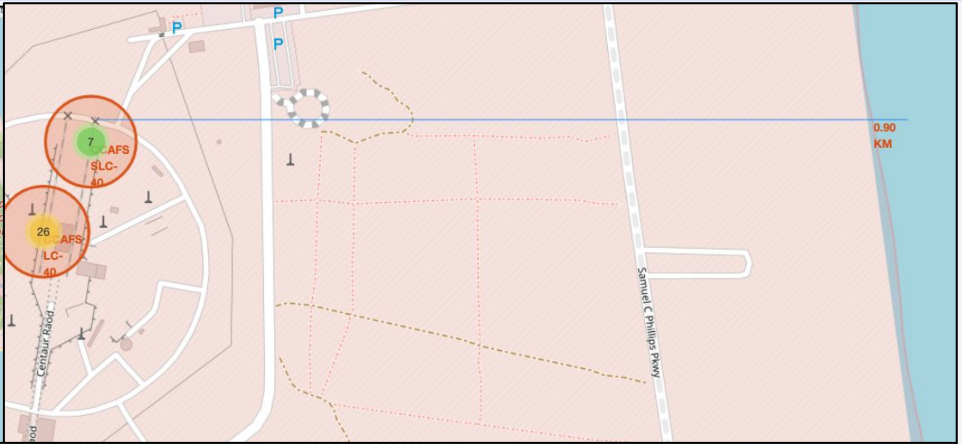
Success/Failed launches marked on Map

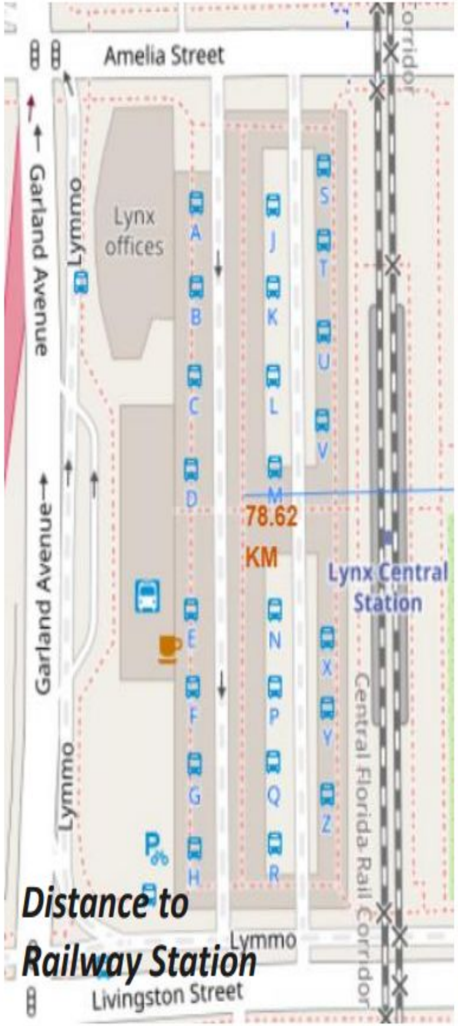
CCAFS SLC-40 CCAFS LC-40 VAFB SLC-4E KSC LC-39A

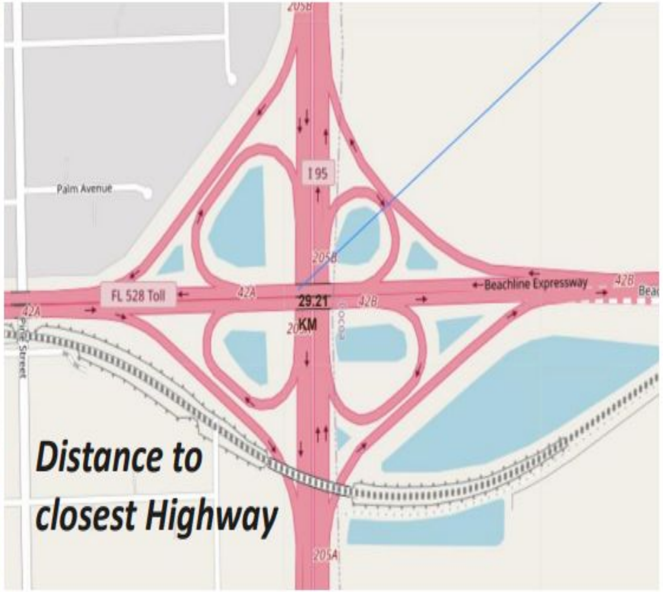
Red Marker Indicates Failure and Green Indicates Success.

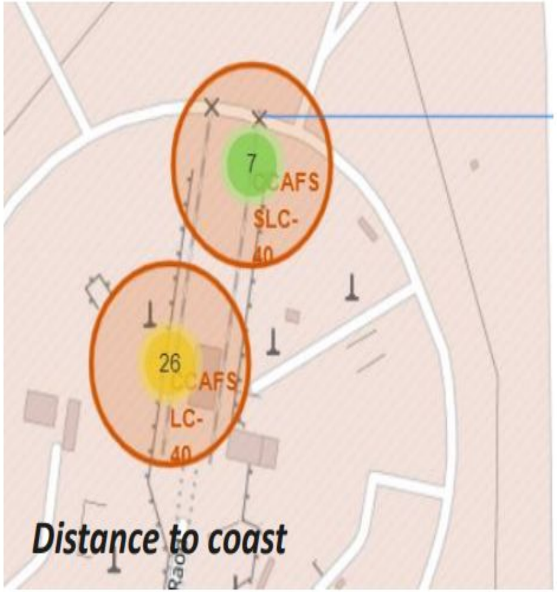
Distance between a launch sites to its Proximities

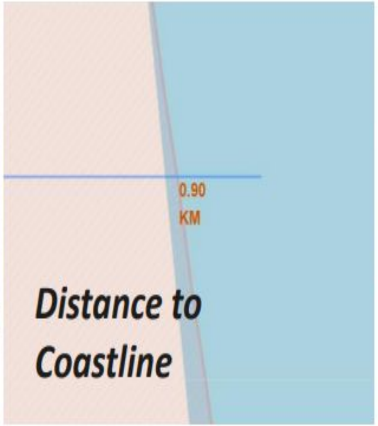


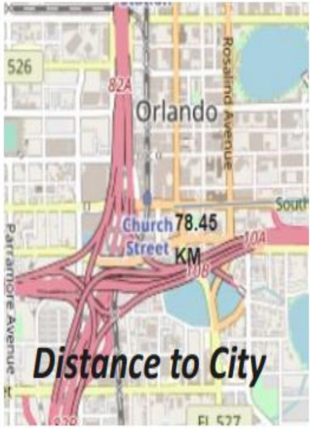










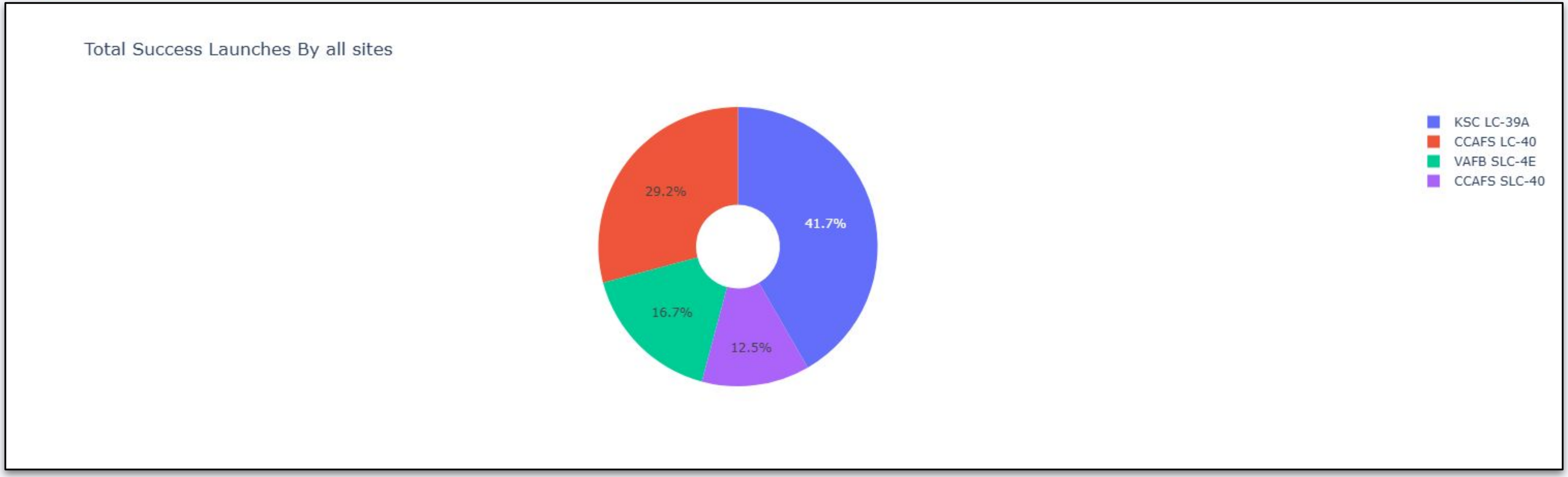


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 5

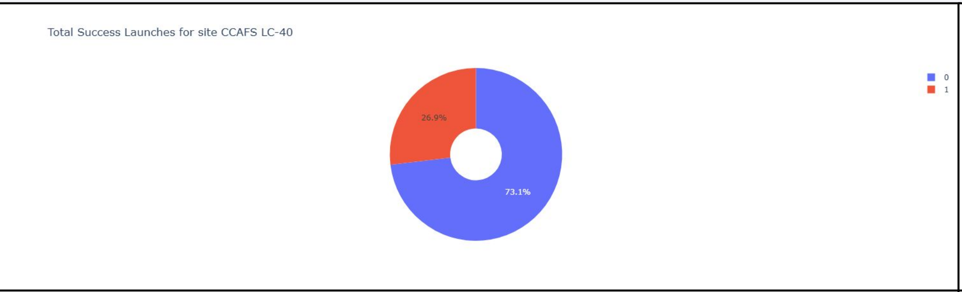
Build a Dashboard with Plotly Dash

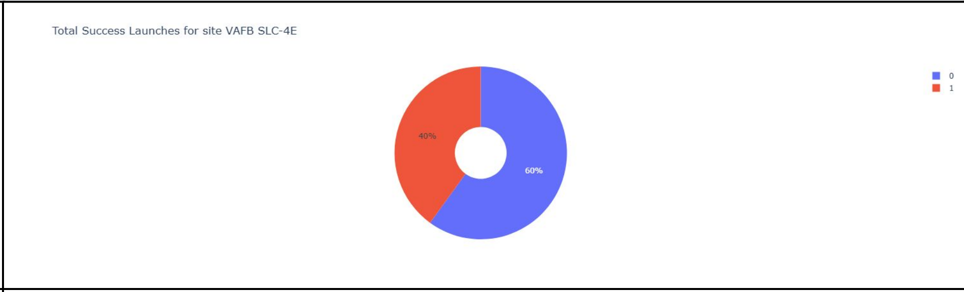
Total Launch Success by all Sites

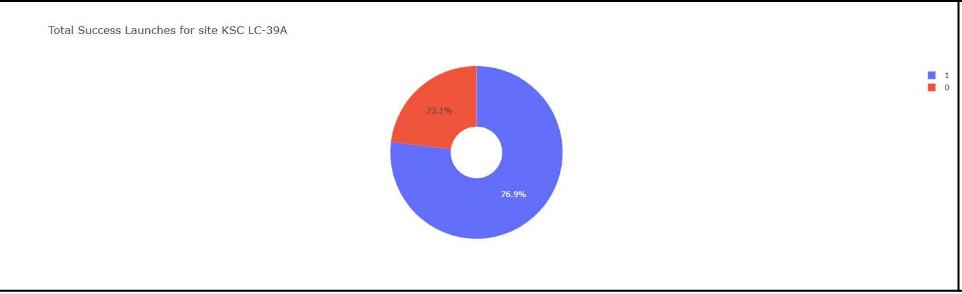


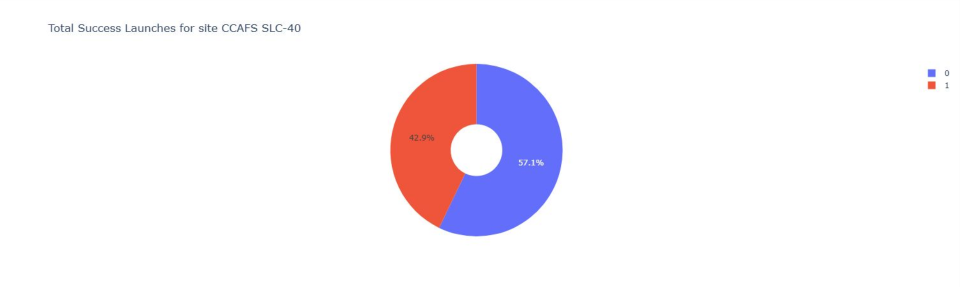
Dash app Code :-
https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex_dash_app.py

Success Rate by Site



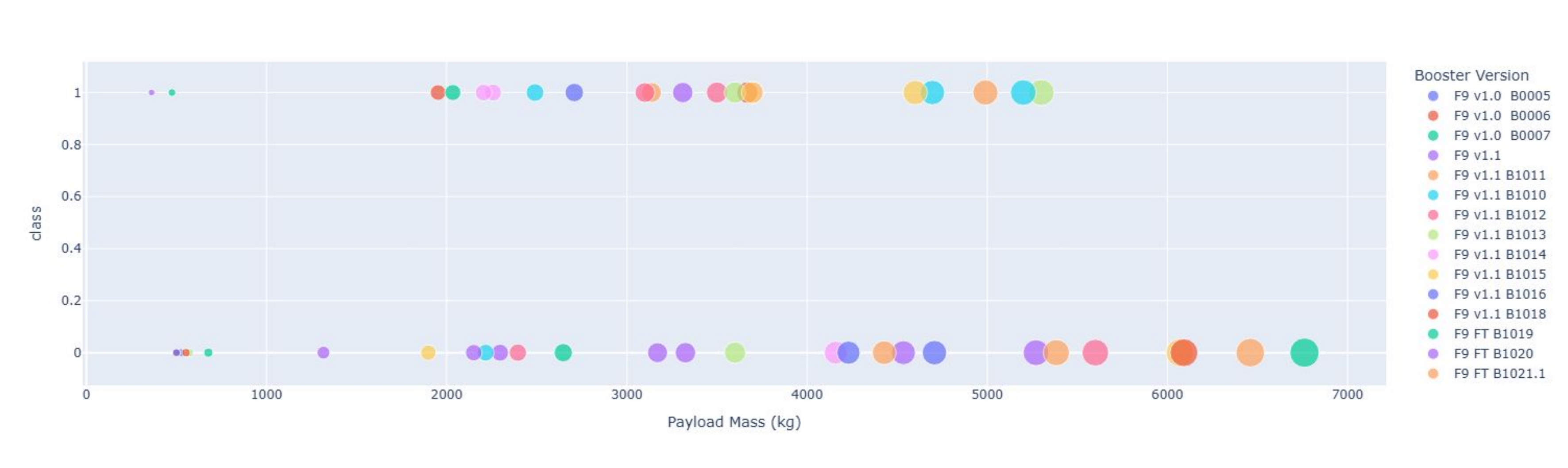






We can see that Kennedy Space Centre KSC LC-39A achieved a 76.9% success rate and had most successful launches of all sites.

Payload vs Launch Outcome



We can see the success rates for low weighted payloads is higher than the heavy weighted payloads.

Section 6

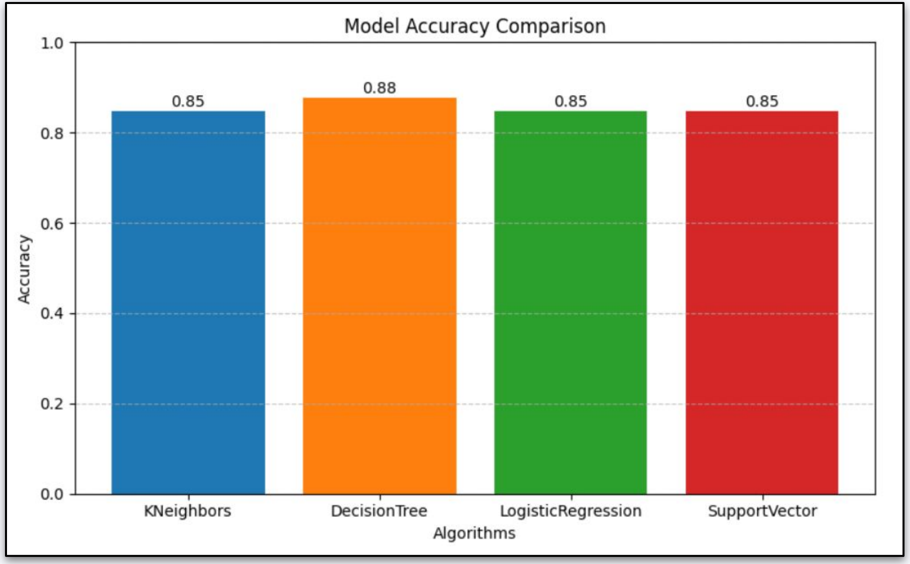
Predictive Analysis (Classification)

Classification Accuracy

Algorithm Accuracy Scores:

- KNeighbors: 0.84
- Decision Tree: 0.87
- Logistic Regression: 0.84
- Support Vector: 0.84

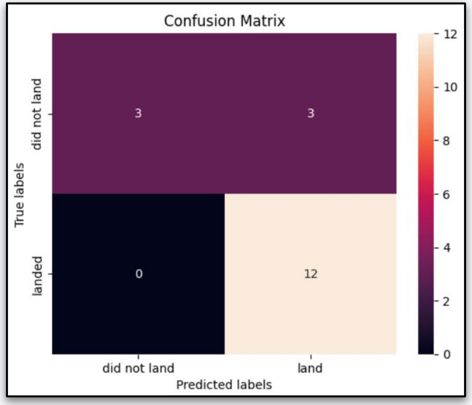
https://github.com/Keshavmishra-hub/SpaceX-DataScience-Capstone/blob/main/spacex_machine_learning_prediction.ipynb

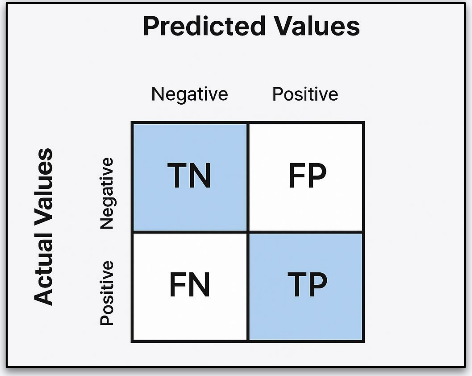



```
Best model is DecisionTree with a score of 0.8767857142857143
Best params is : {'criterion': 'gini', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}
```

After selecting the best hyperparameters for the decision tree classifier using the validation data, we achieved 83.33% accuracy on the test data.

Confusion Matrix





📊 **Key Insights:**

1. **High True Positive Rate (TP = 12):** The model correctly predicted **all 12 landings**—this shows excellent accuracy for identifying successful landings.

2. **Moderate False Positives (FP = 3):** The model predicted 3 rockets **would land**, but they **did not**. This suggests the model is slightly overconfident in predicting successful landings.

3. **Zero False Negatives (FN = 0):** Importantly, **no landed rockets were missed**. This is crucial in high-stakes scenarios where missing a success would be costlier than a false alert.

4. **Balanced on Non-landings (TN = 3):** The model also correctly identified 3 out of 6 non-landings, indicating a **50% accuracy for failed landings**.

✅ **Conclusion:** The Decision Tree model is **very good at detecting successful landings** (recall = 1.0 for class "landed") but has **moderate precision**, meaning it sometimes predicts a landing even when one doesn't happen. This makes it useful when minimizing missed landings is more important than occasionally raising a false alert.

Conclusions

- The Decision Tree Classifier Algorithm is the best for Machine Learning for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- We can see that KSC LC-39A had the most successful launches from all the sites.
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate.

Thank you!