IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Zin Thu Thu Latt>
<12 August 2025>

Outline

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

Executive Summary

Summary of methodologies

1. Data Collection
    a. Access SpaceX launch data from API and web scrapped
2. Data cleaning & Preparation
    a. Data cleansing
    b. Store data
    c. Perform SQL queries
    d. Conduct EDA
3. Create New and Standardized data
4. Interactive Visualization
    a. Ma and success rate using Folium
    b. Interactive Board with Plotly Dash
5. Model Building and Evaluation
    a. Implemented SVM, Decision Tree and K-nearest Neighbours
    b. Test data accuracy

Summary of all results

1. Data Insights
    a. Identified Falcon 9 landings
    b. Visualized geographical patterns and success rates
2. Model Performance
    a. SVM and K-nearest with 83.3% accuracy
    b. Decision tree with 94.44% accuracy
3. Key Findings
    a. Launch site and payload mass impact landing success
    b. Decision tree being the most effective predictor

Introduction

Project background and context

This project is based on SpaceX Falcon 9 rocket. With a high cost of launch spaceship, by using the public data and information, we get to know about the reusability and the cost of the next launch. These information are retrieved and analyzed by machine learning models.

Problems you want to find answers

1. What impacts the success rate of the landing?
2. Predicting the landing outcome using machine learning models
3. Which machine learning model performs best in predicting the success landing rate?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology

- Sourced from SpaceX RestAPI, from web scraping from Wikipedia

Perform data wrangling

- Data cleaning includes missing values, filtering data by using one hot encoding to prepare data to a binary classification

Methodology

Executive Summary

Perform exploratory data analysis (EDA) using visualization and SQL

- Visualization displaying success rates and payloads using Matplotlib and Seaborn
- Executing SQL queries to get objective questions

Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models
- Develop Plotly Dash with interactive components like dropdowns and sliders to analyze success rate and payload ranges

Data Collection

API requests from API and Web Scrape

API Request → Fetch and Launch data → Store Data Locally

Web Scrape

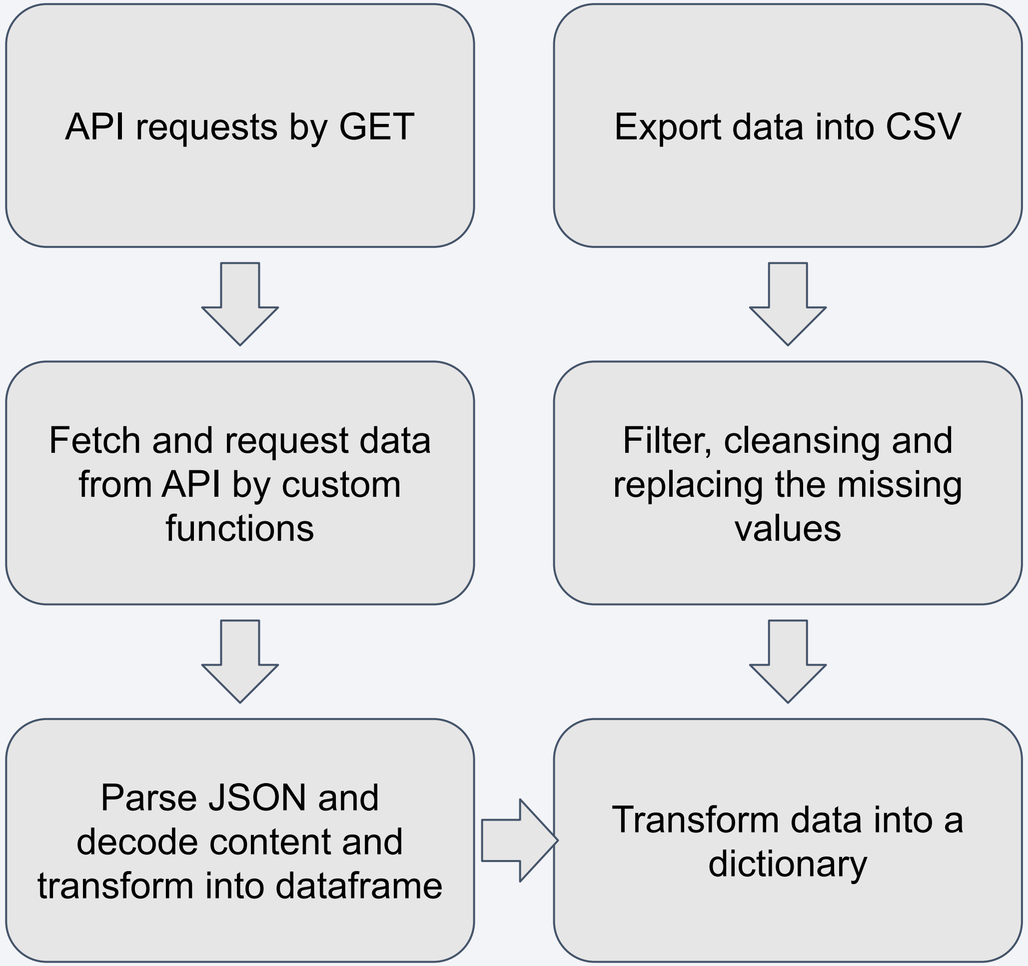
Extract HTML → Parse with Beautifulsoup → Convert to DataFrame

Data Collection – SpaceX API

From both SpaceX RestAPI and Web scrapping from Wikipedia to get a complete dataset for this project.

Source link by RestAPI: https://api.spacexdata.com/v4/launches:

Github link:



Data Wrangling

During this project, a number of cases where booster did not land successfully and sometimes, a landing was attempted but it was due to an accident.

True Ocean - successfully landed to an ocean

False Ocean - unsuccessful landed to a specific part of the ocean

True RTLs - successfully landed to a ground pad

False RTLs - unsuccessfully landed to a ground pad

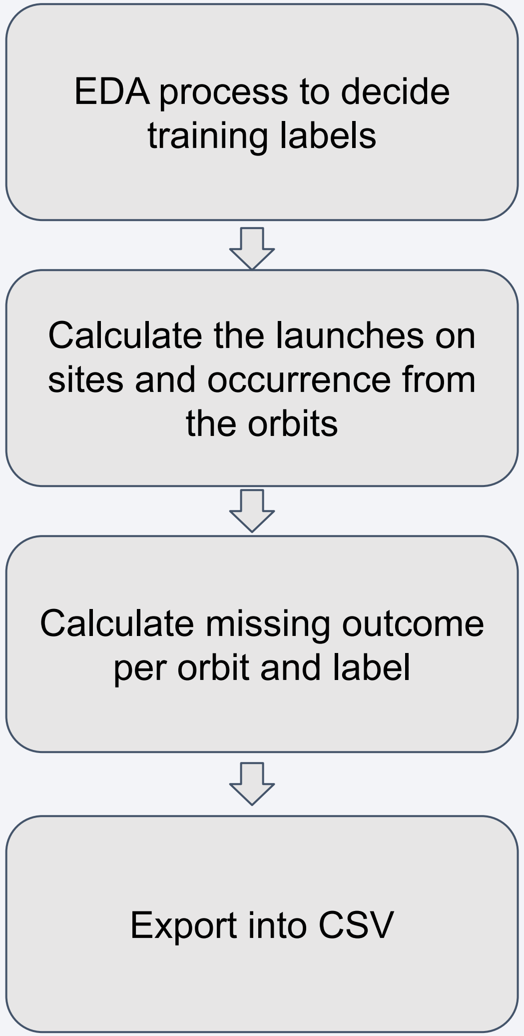
True ASDS - successfully landed on a drone ship

False ASDS - unsuccessfully landed on a drone ship

From the above labels, transforming into,

1 - successful landed booster

0 - unsuccessful landed booster



EDA with Data Visualization

Summarized the overall state of the dataset and display them visually.

1. Bar Chart

   Displays the comparisons among categories and produce a measured value. Like success and fail launches and rocket types

2. Line Chart

   Displays trends over time and the overall time series of the successful rate of Falcon 9 launches

3. Scatter Plot

   Displays relationships between a variables and if relationships exist, they can be used for ML models. Like payload mass vs successful launches and how one variables affects and concerns one another

EDA with SQL

- Calculate total number of success and fail launches
- Calculate success rate based on launch sites and rocket types
- Join tables and datasets for comprehensive analysis
- Sorted data to witness trends and outliers
- Displayed 5 records of launches being with string CCA
- Witness average payload mass carried by booster launched by NASA and F9
- Rank the landing outcomes, that is, failure drone ship or success ground pad between 2010 and 2017

GitHub URL

Build an Interactive Map with Folium

Markers

To indicate launch sites on the map, and each represent a specific geographical location where launches occurred. Pinpointing each launching location for spatial reference.

Circles

Visual appeal to represent launch sites and witness the potential impact zones around launch sites. Also helps to know the safety perimeters and operational boundaries

Lines

Visual appear to display the relevant locations, spatial context and connection between different points of interest.

- GitHub URL

Build a Dashboard with Plotly Dash

Pie Chart

Display success and failed launches and visualize the overall success rate and performance trends.

Scatter Plot

Display the relationship between payload mass and launch success in order to explore how it influences mission outcomes

Dropdown

Display specific launch sites to analysis, filter and focused on exploration based on geographical locations

Slider

Adjustable slider to display payload mass to exanime successful launches

- GitHub URL

Predictive Analysis (Classification)

Data processing

- Equal variables
- Split into training and testing sets in 80/20 ratio

Model Selection

- SVM, decision tree, KNN which are suitable for binary classification

Hyperparameter Tuning

- Used GridSearchCV to systematically search for optimal hyperparameters
- Tuned parameters like SVM, max depth and n_neighbouts

Model Evaluation

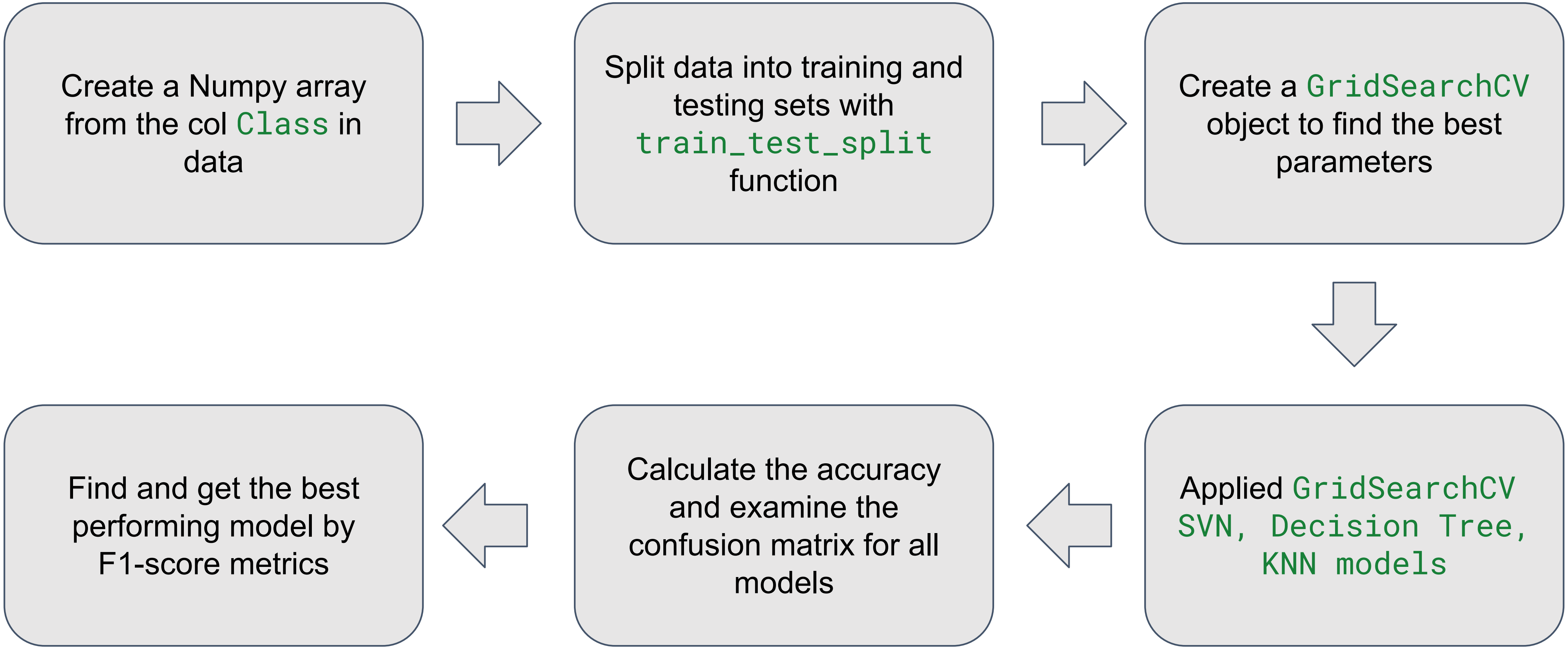
- Cross-validation technique to ensure robustness and generalizability
- Accuracy, precision, recall and F1-score to access model performance

Improvement Iterations

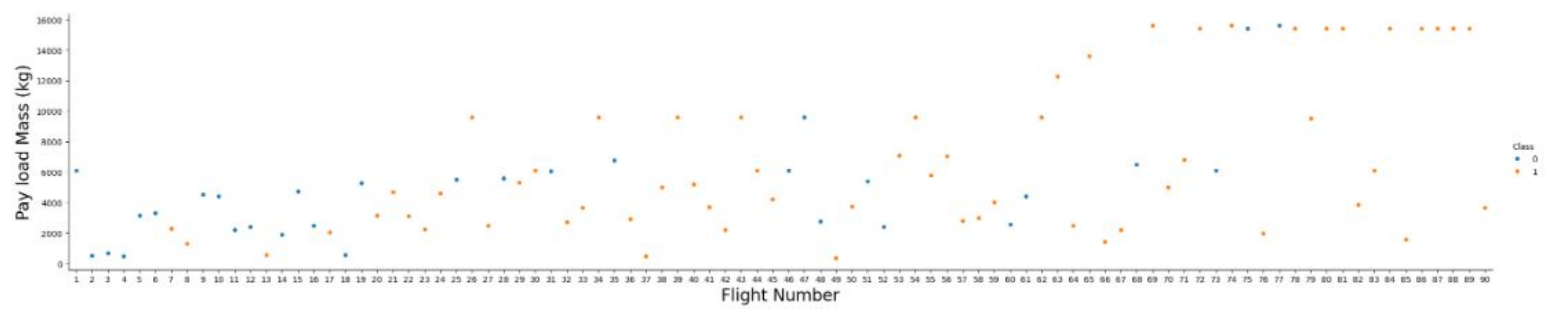
- Fine-tuning hyperparameters to maximize predictive accuracy and reliability

GitHub URL

Predictive Analysis (Classification)

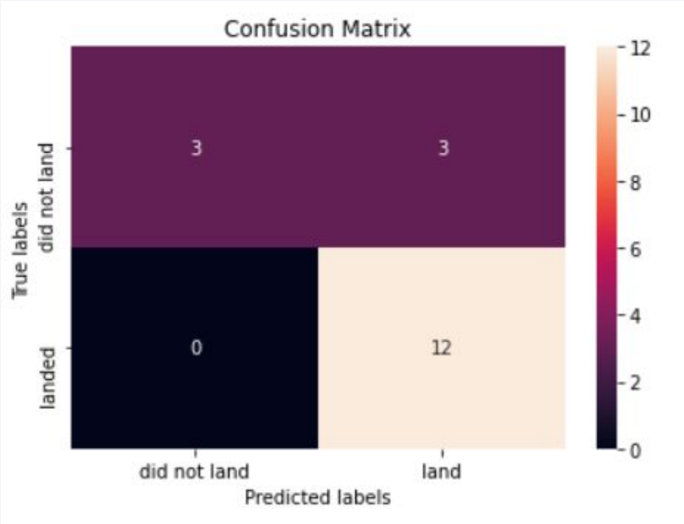


Results



EDA Result

Interactive Analytics

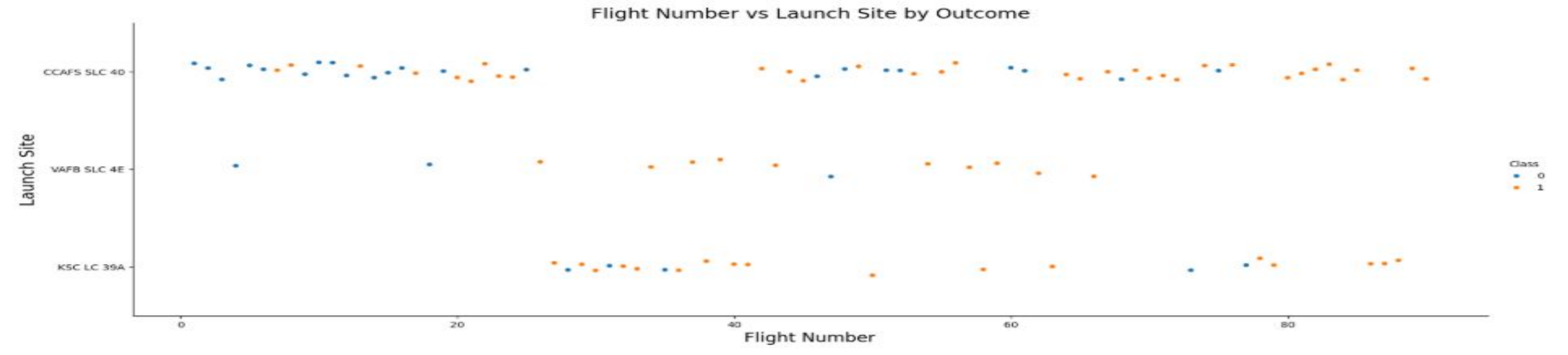


Predictive analysis results

Section 2

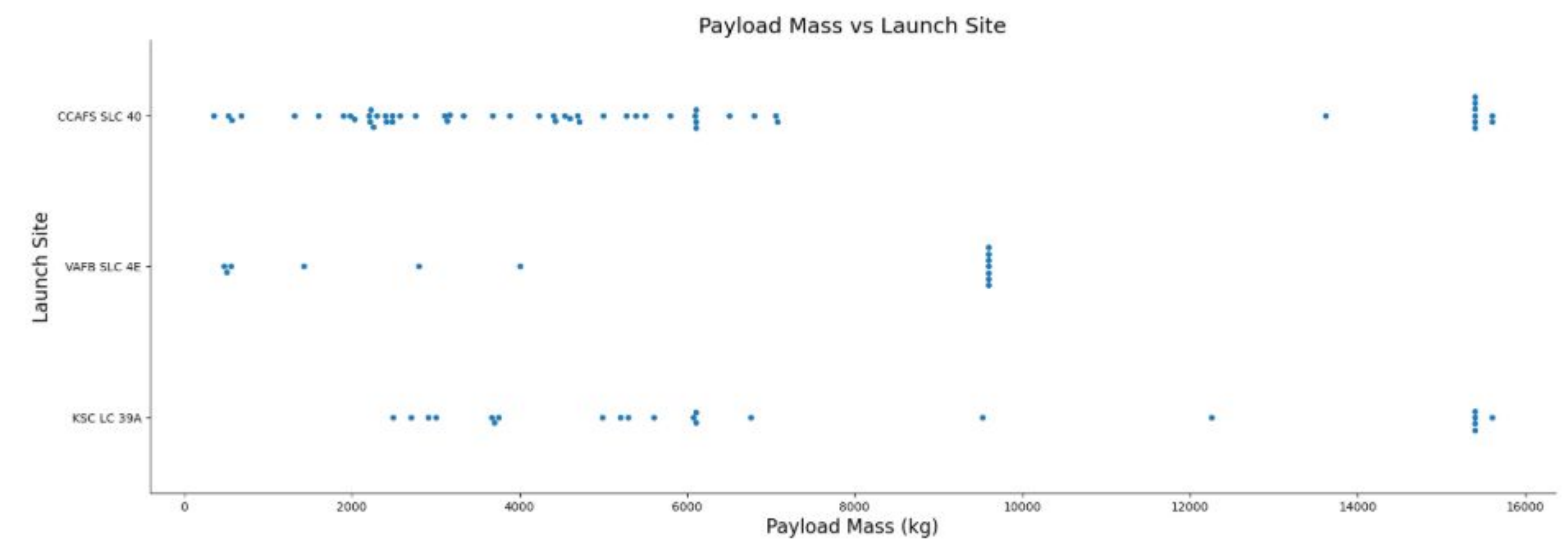
Insights drawn from EDA

Flight Number vs. Launch Site



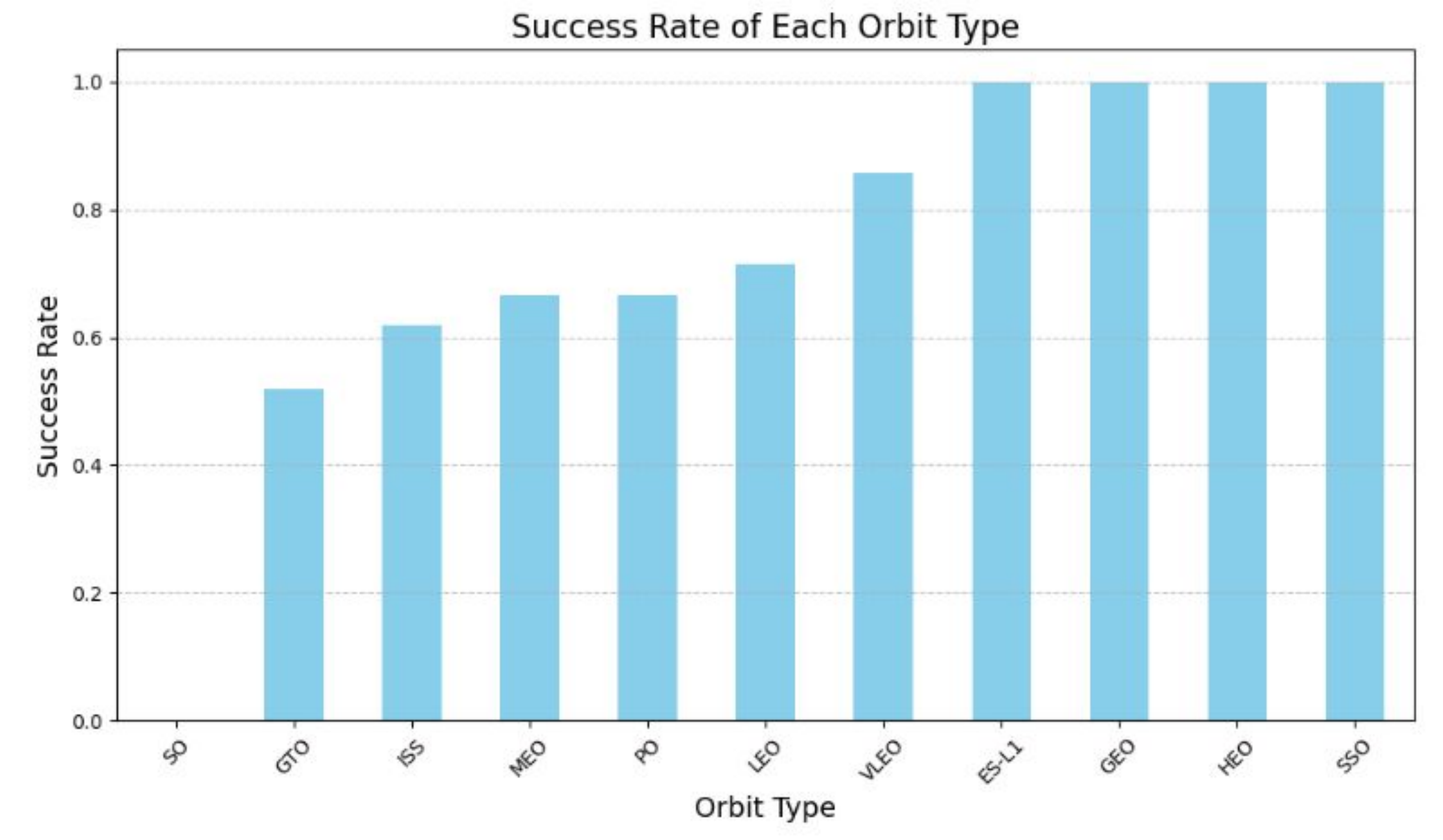
- CCAFS SLC 40 and KSC LC 39A have a mix of success and unsuccess landings
- CCAFS SLC has the most number of launches which is about half of total launches
- VAFB SLC 4E and KSC LC 39A have higher success landings
- Summary, earliest and new launches has a higher rate of success

Payload vs. Launch Site



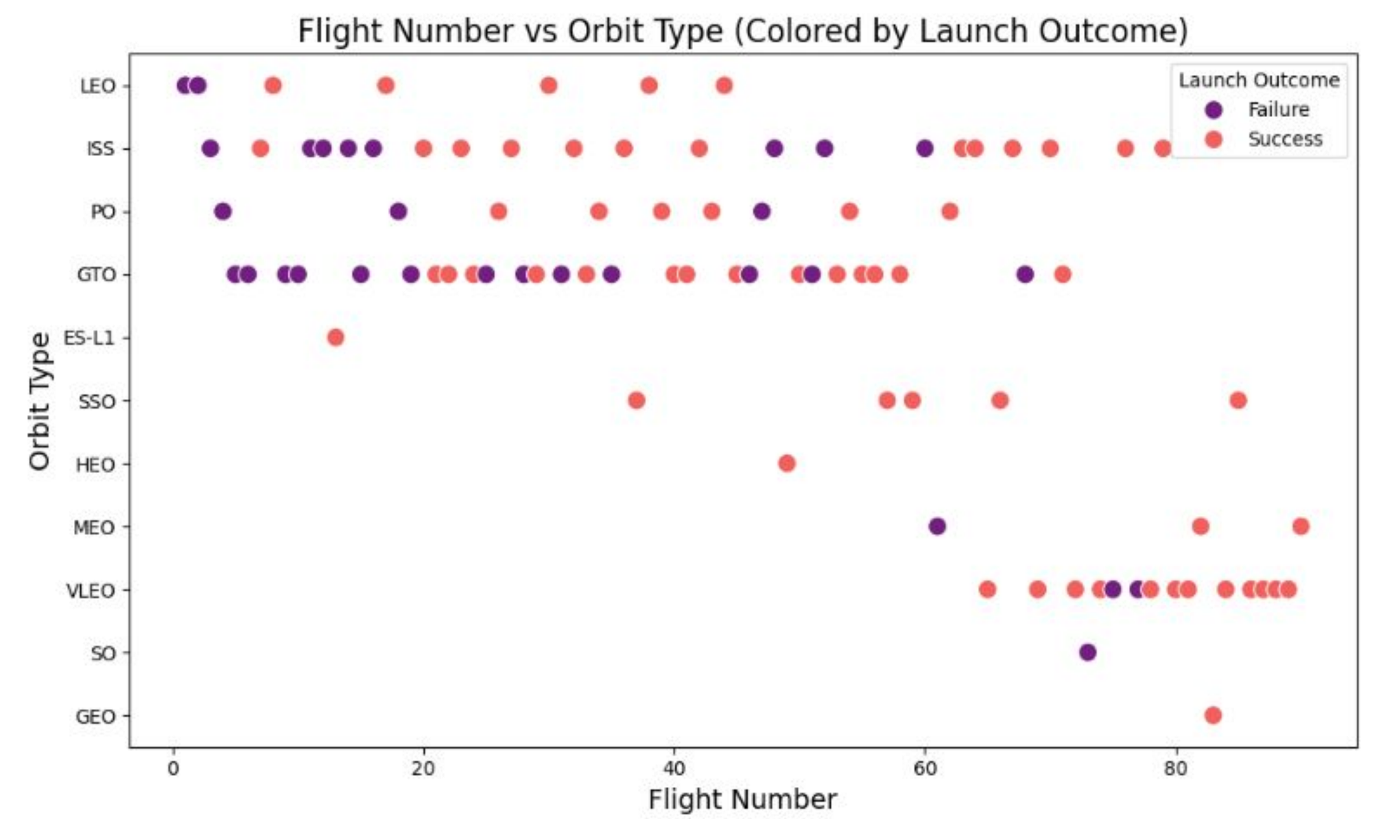
- CCAFS SLC 40 site mostly handle payloads below 10,000 kg
- VAFB SLC 4E and KSC LC 39A sites handle a wider range of payload
- KSC LC 39A is more frequently used for heavier payloads over 15,000 kg and suitable for high-capacity missions. Also has a 100% success rate for payload mass under 5500 kg
- Higher the payload mass, higher the launch site success tate

Success Rate vs. Orbit Type



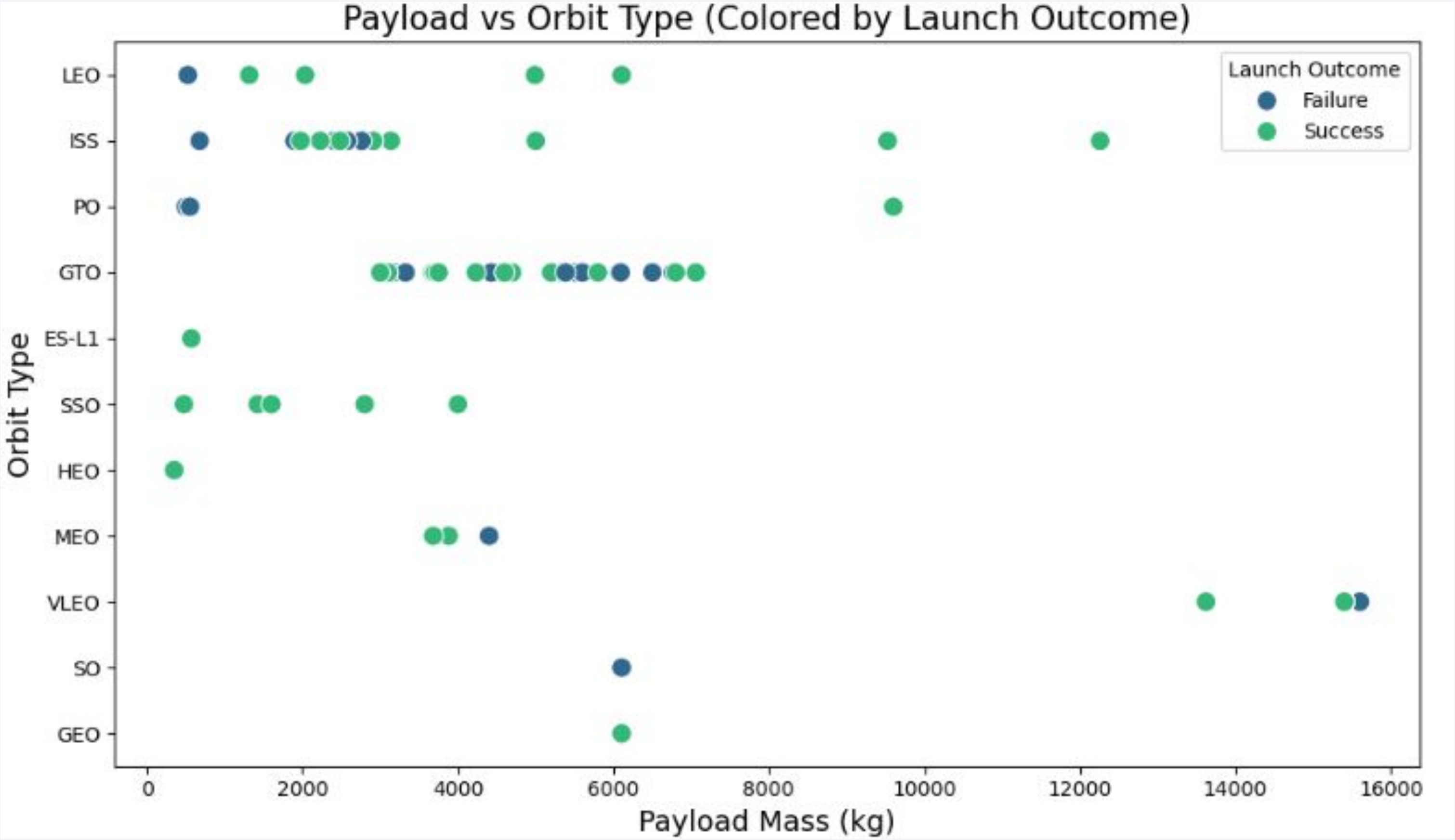
- Orbits VLEO, ES-L1, GEO, HEO and SSO have 100% success rate and can assume that these orbits are reliable for first stage landings
- GTO with half success rate can be assumed to have complexity and/or challenges with landing.
- SO orbit has 0% success rate

Flight Number vs. Orbit Type



- Falcon 9 improves eventually with higher flight number
- GTO and ISS have a mixed of success launch outcomes and also get better in the recent mission

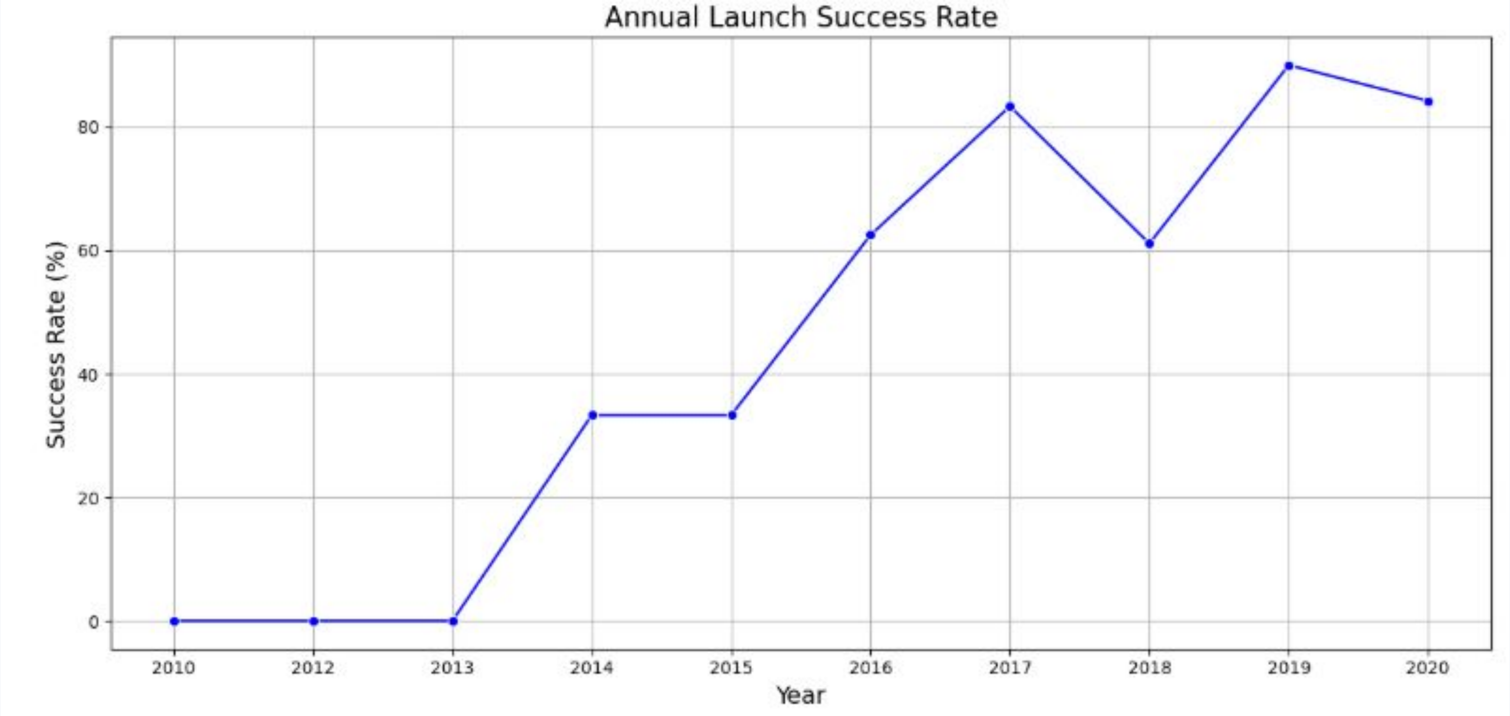
Payload vs. Orbit Type



- Payloads under 6000 kg have more successful landings
- Above 10,000 kg payloads have a mix of success and fail landings
- Heavier the payloads, the higher the difficulty in landing

Launch Success Yearly Trend

- Significant increase in success launching rate after 2013 with over 80% success rate in 2020
- With a dip in 2018 but overall improvements over the years in Falcon 9 launches



All Launch Site Names

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

Names of unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

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

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [30]:

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

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

Out[30]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

In [34]:

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1
```

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

Out[34]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [36]:

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pa
```

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

Out[36]:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [38]:

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Suc
```

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

Out[38]:

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [40]:

```
%sql SELECT "Mission_Outcome", COUNT(*) AS "Total" FROM SPACEXTABLE WHERE "Mission_Out
```

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

Out[40]:

Mission_Outcome	Total
Success	98

Boosters Carried Maximum Payload

```
Done.
Out[42]:
```

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

```
Done.
Out[69]:
```

Month_Name	Mission_Outcome	Booster_Version	Launch_Site
January	Success	F9 v1.1 B1012	CCAFS LC-40
February	Success	F9 v1.1 B1013	CCAFS LC-40
March	Success	F9 v1.1 B1014	CCAFS LC-40
April	Success	F9 v1.1 B1015	CCAFS LC-40
April	Success	F9 v1.1 B1016	CCAFS LC-40
June	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
December	Success	F9 FT B1019	CCAFS LC-40

List of the failed `landing_outcomes` in drone ship, their booster versions, and launch site names for in year 2015

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

```
    "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    "Landing_Outcome"
ORDER BY
    COUNT(*) DESC;
```

```
 * sqlite:///my_data1.db
Done.
Out[81]:
```

Landing_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

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, in descending order

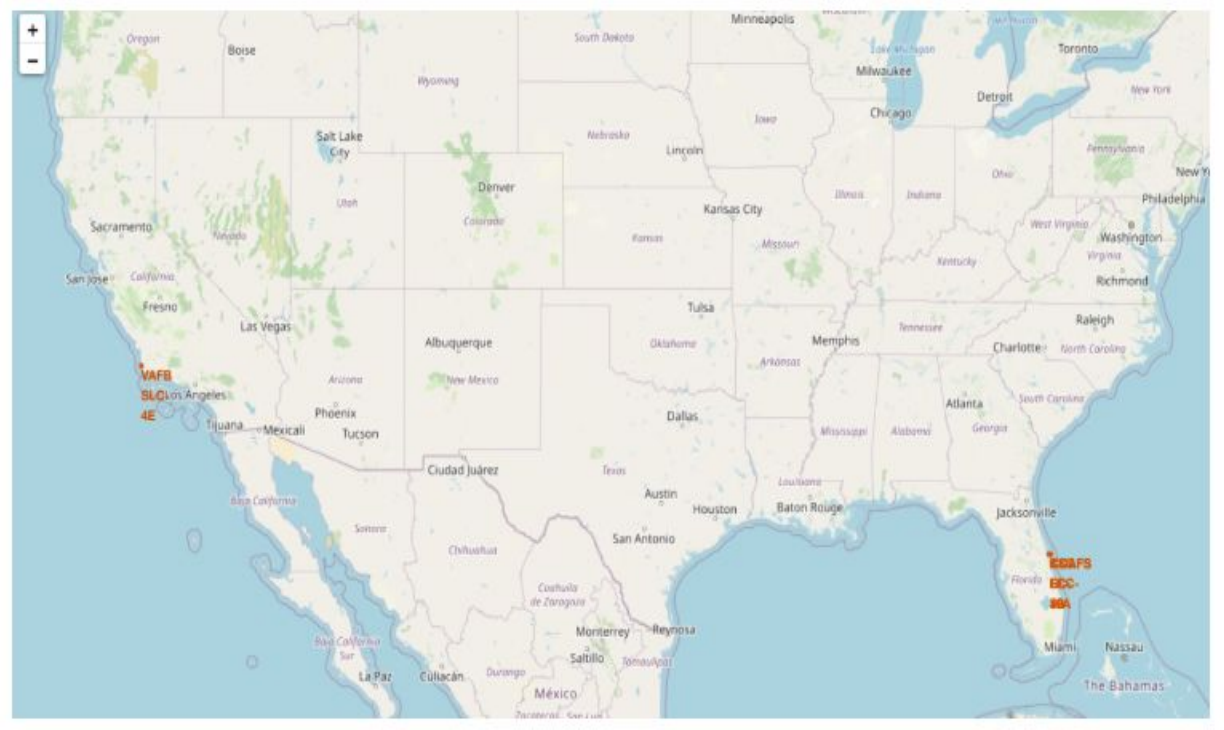
Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- Most launches are closed to equator line since the land moves faster at equator than any other place on the Earth. Only Vandenberg Air Force Base happened to be in Florida at a latitude of 34.63

- All launch sites are close to the coast, since it reduce the risk of having any debris dropping or exploding near big cities.

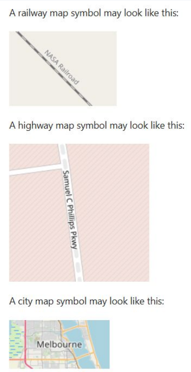


<Folium Map Screenshot 2>

- Visualization with clustered markets allows to witness a clearer SpaceX launch sites and make it easier to manage a larger number of markers to observe patterns
- The red markers, unsuccessful launches and the green markers, successful launches displays immediate visual feedback on the performance of launches at each sites
- KSC LC-39A launch sites has a very high success rate

<Folium Map Screenshot 3>

- Visual representation represents the distance between a launch site to its proximities, that is, CCAFS SLC-40 and its closest coastline
- Calculated distance is approximately 0.51 kilometers
- Additional PolyLine displays the straight line distance and highlighting the proximity of the launch site to the coast
- KSC LC-39A is close to its closest city Titusville, 16.32 km



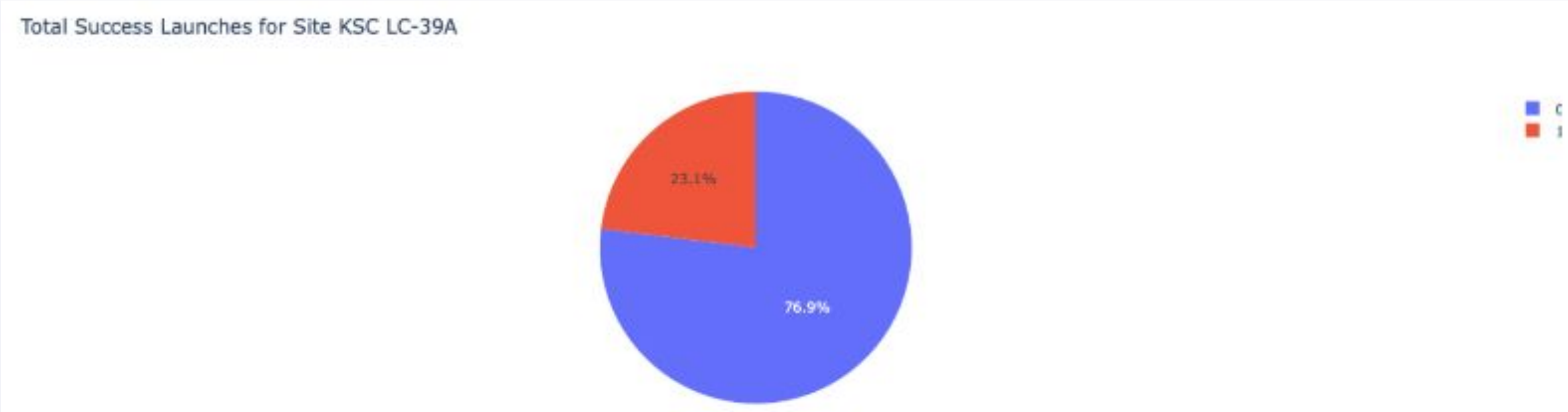
Section 4

Build a Dashboard with Plotly Dash

Count for all successful launch sites

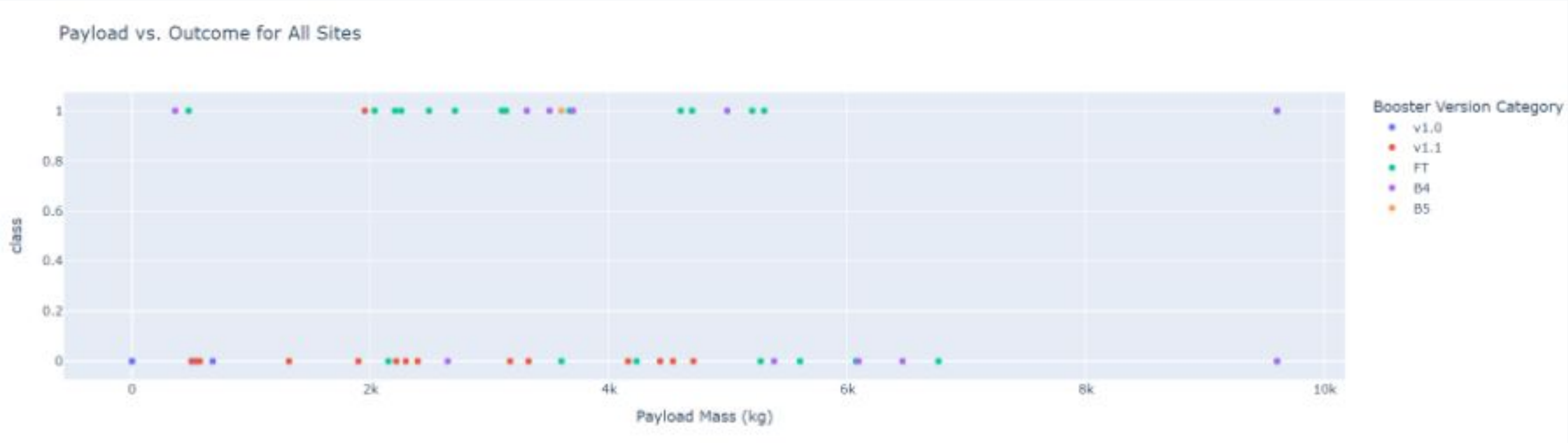
Displaying launch success count for all sites, in a piechart

Highest success rate launch sites



Pie Chart for the launch site with highest launch success ratio where KSC LC-39A has the highest launch rate with 76.9% with 10 successful landings and only 3 failed attempts

Payload vs. Launch Outcome



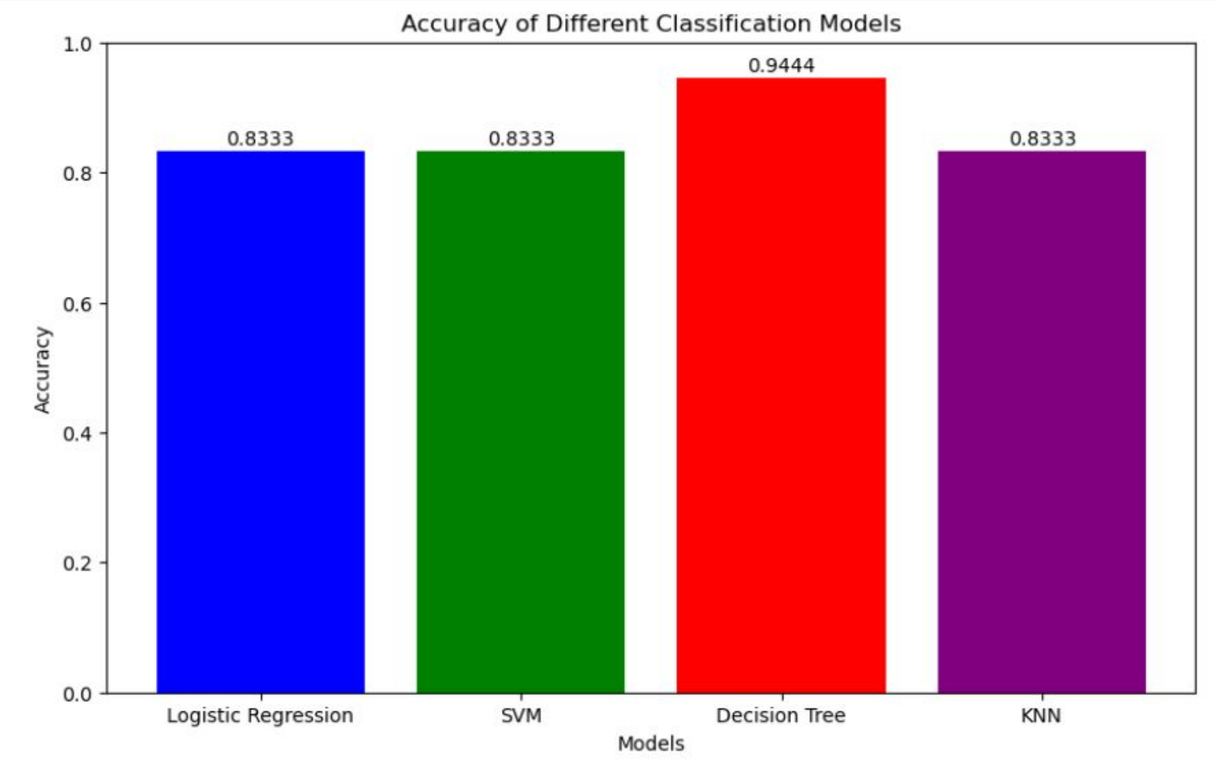
Payload vs. Launch Outcome in scatter plot for all sites, with different payload selected in the range slider which shows payloads between 2000 and 5500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

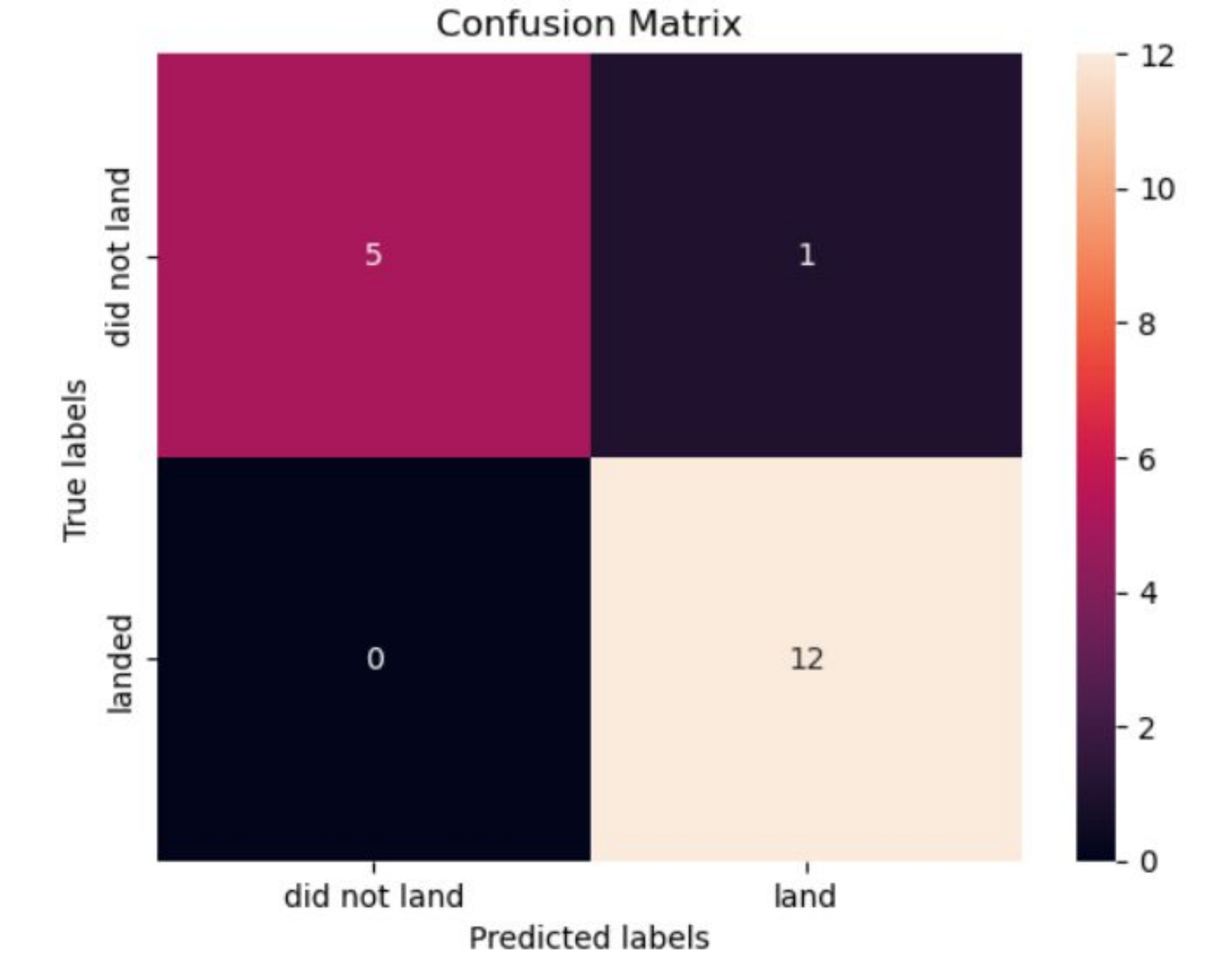
Classification Accuracy

- Decision tree has the highest classification rate with 0.944. Not only high in score but also with the highest accuracy
- The rest of the classification with the accuracy of 0.8333



Confusion Matrix

- Logistic regression can distinguish between different classes and the major problem is false positives
- High accuracy score of 94.44% with a significant number of true positive and true negatives, displaying the effectiveness in predicting Falcon 9 first stage landings



Conclusions

- Decision tree is the best model for this case study
- CCAFS LC-40 launch site has the highest success rate with 43.7% successful launches
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- Overall success rate of launch sites increases over the years
- Launches with low payload mass has higher success rates than launches with high payloads
- Most launches happen near the Equator line and very close to the coast in order to avoid unnecessary damages

Thank you!