Winning Space Race with Data Science

Vladimir Kuridza

15/06/24

Outline

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

Executive Summary

Objective: Predict SpaceX rocket landing success and identify key influencing features using machine learning.

Methodology:

1. **Data Collection:**
    1. Web scraped historical SpaceX launch data, including launch conditions, rocket types, and landing outcomes.
2. **Data Preparation:**
    1. Cleaned and pre-processed data into a structured data frame, handling missing values and encoding categorical variables.
3. **Exploratory Data Analysis:**
    1. Applied visualizations to study data distributions, patterns, and correlations.
4. **Feature Engineering:**
    1. Created new features to enhance model predictions.
5. **Model Building:**
    1. Applied machine learning techniques (logistic regression, decision trees, random forests) to predict landing success.
    2. Achieved over 80% accuracy in predicting rocket landing outcomes.
6. **Feature Importance Analysis:**
    1. Identified key features influencing landing success using feature importance scores and SHAP values.

Results: Machine learning models can predict SpaceX rocket landing success with over 80% accuracy, providing valuable insights into the factors that impact landing outcomes. These insights can help improve launch strategies and landing success rates.

Introduction

- SpaceX, a leader in aerospace innovation, has revolutionized the space industry with its reusable rocket technology. The ability to successfully land and reuse rockets not only reduces costs but also increases the frequency of missions. However, achieving consistent landing success is complex and influenced by various factors. This project aims to leverage machine learning techniques to predict whether a SpaceX rocket will land successfully and to identify the key features that influence these outcomes.
- By the end of this project, we aim to develop a reliable predictive model for SpaceX rocket landings and gain insights into the factors that most significantly influence landing success. These insights could help SpaceX enhance their launch strategies, improve landing success rates, and further advance the goal of cost-effective space exploration.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using a collection of API calls and Web scraping commands
- Perform data wrangling
  - We identified any missing data and replaced it with mean values. We also created an extra column for the outcome of each landing.
- 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

Overview of data extraction methods:

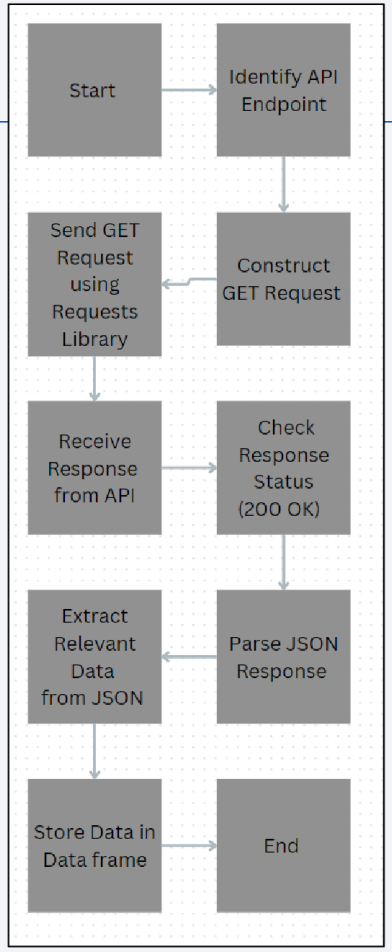
- Web Scraping: Extracting data from websites by parsing HTML.
- API Calls: Extracting data from web services using predefined endpoints.

Tools Used:

- Beautiful Soup: A Python library for web scraping.
- Requests: A Python library for sending HTTP requests.
- APIs: Application Programming Interfaces provided by web services.

Data Collection – SpaceX API

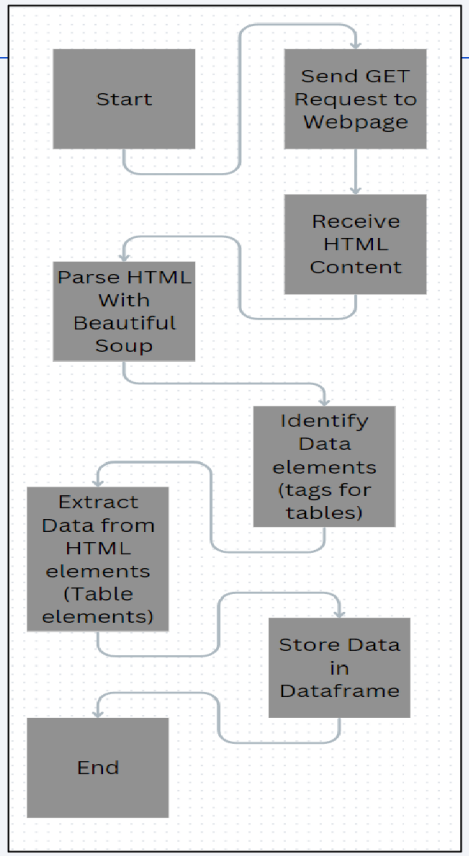
- Data collection with SpaceX REST calls process:
- https://github.com/ArthurMan100/Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

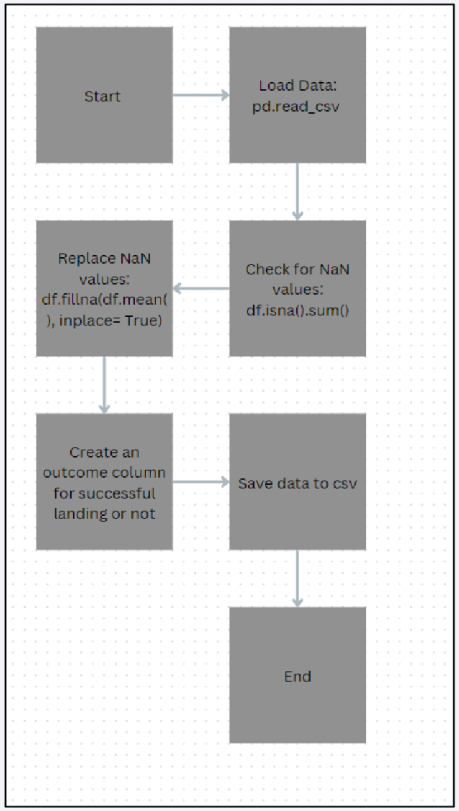
- Flow Chart for Web scraping techniques used

https://github.com/ArthurMan100/Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- https://github.com/ArthurMan100/Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter graphs, bar graphs and line graphs were used for the exploratory data analysis stage.
- Scatter charts are good for identifying and illustrating relationships between two numerical variables. They can help to see if there is a correlation, pattern, or outliers.
- Categorical Data Visualization: Bar graphs are ideal for comparing categorical data. They make it easy to see which categories are the most or least frequent.
- Line graphs are perfect for visualizing trends over time. They help in understanding how a variable changes over a continuous period.
- https://github.com/ArthurMan100/Capstone-Project/blob/main/edadataviz.ipynb

EDA with SQL

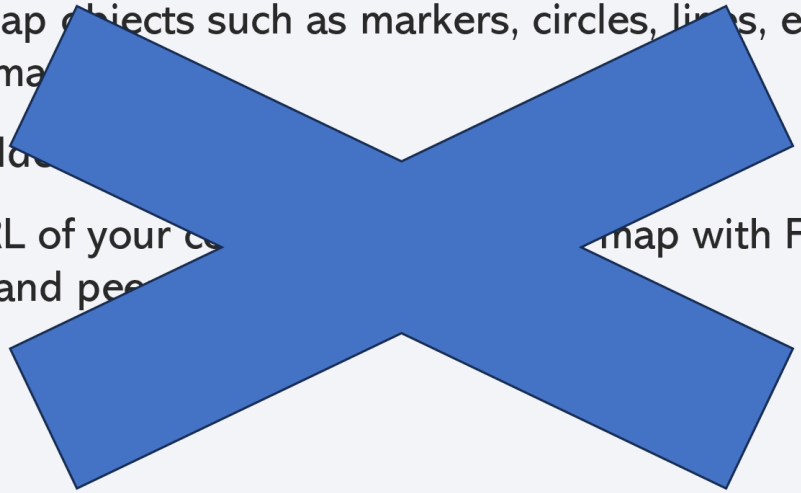
Summary of SQL Queries Performed as part of the EDA stage

- Ranked Count of Landing Outcomes Between 2010-06-04 and 2017-03-20
- Names of Boosters that Carried the Maximum Payload Mass
- Count of Launch Outcomes
- Date of the First Successful Landing on a Ground Pad
- Average Payload Mass for Booster Version F9 v1.1
- First 5 Records Where Launch Sites Begin with CCA
- List of Unique Launch Sites
- Total Payload Carried by NASA Boosters

- https://github.com/ArthurMan100/Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(3).ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer

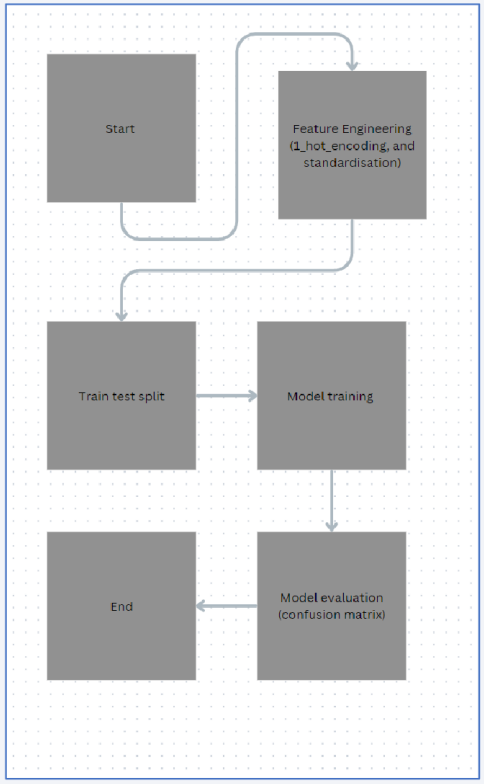


Build a Dashboard with Plotly Dash

- I created a plotly dashboard to visualize successful launches by site using a pie chart, and also being able to filter by payload range (Kg)
- I also plotted a correlation between payload and success rate for all sites using a scatter graph.
- I added a drop-down list of all the sites.
- GitHub link: https://github.com/ArthurMan100/Capstone-Project/blob/main/spacex_dash_app%20(1).py

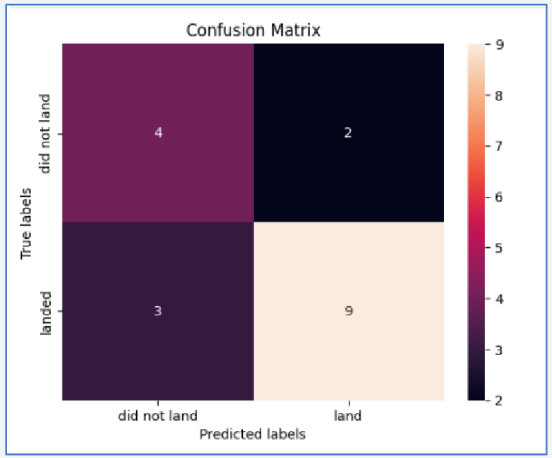
Predictive Analysis (Classification)

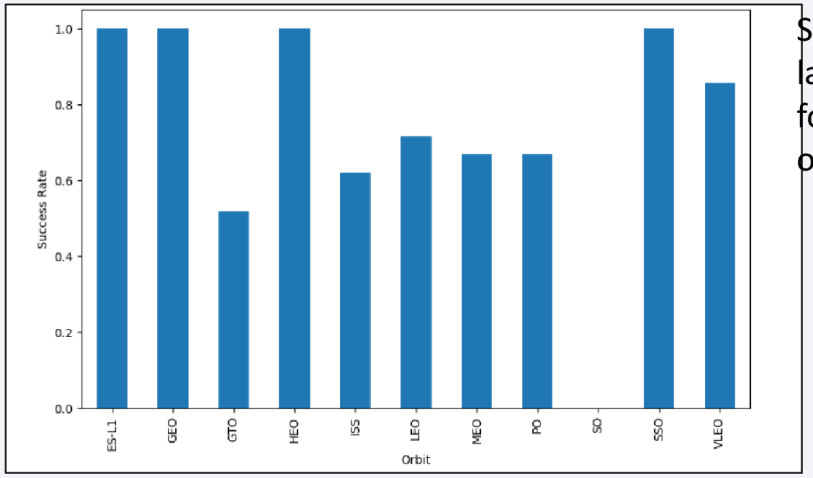
- I first used one_hot_encoding to separate all variables into their own feature columns. I then standardized the data and split the data into training and testing data using train_test_split.
- I then trained the data and built 4 machine learning models. Logistic regression, KNN, Decision trees Classifier and GridSearchCV.
- I determined the best model by calculating the accuracy and plotting confusion matrixes
- https://github.com/ArthurMan100/Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(2).ipynb



Results Summary

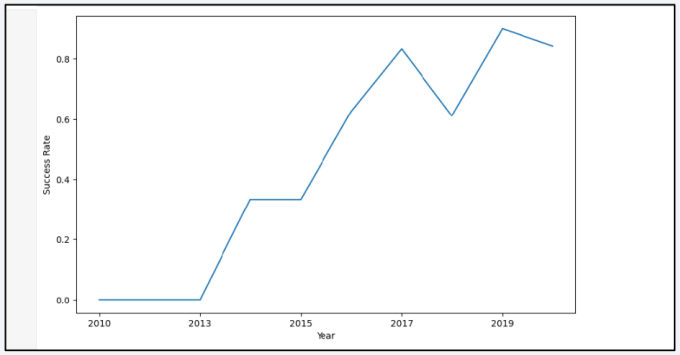
Confusion matrix for classification modelling

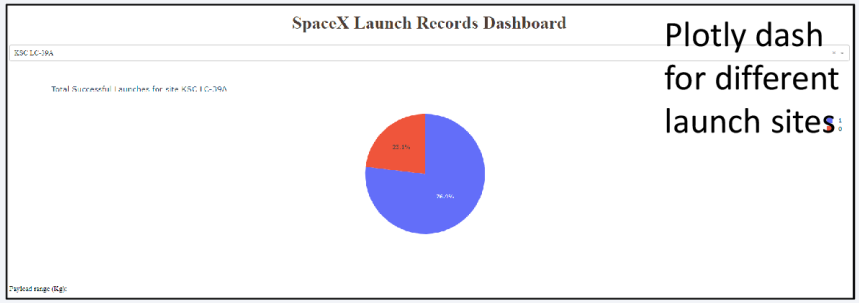




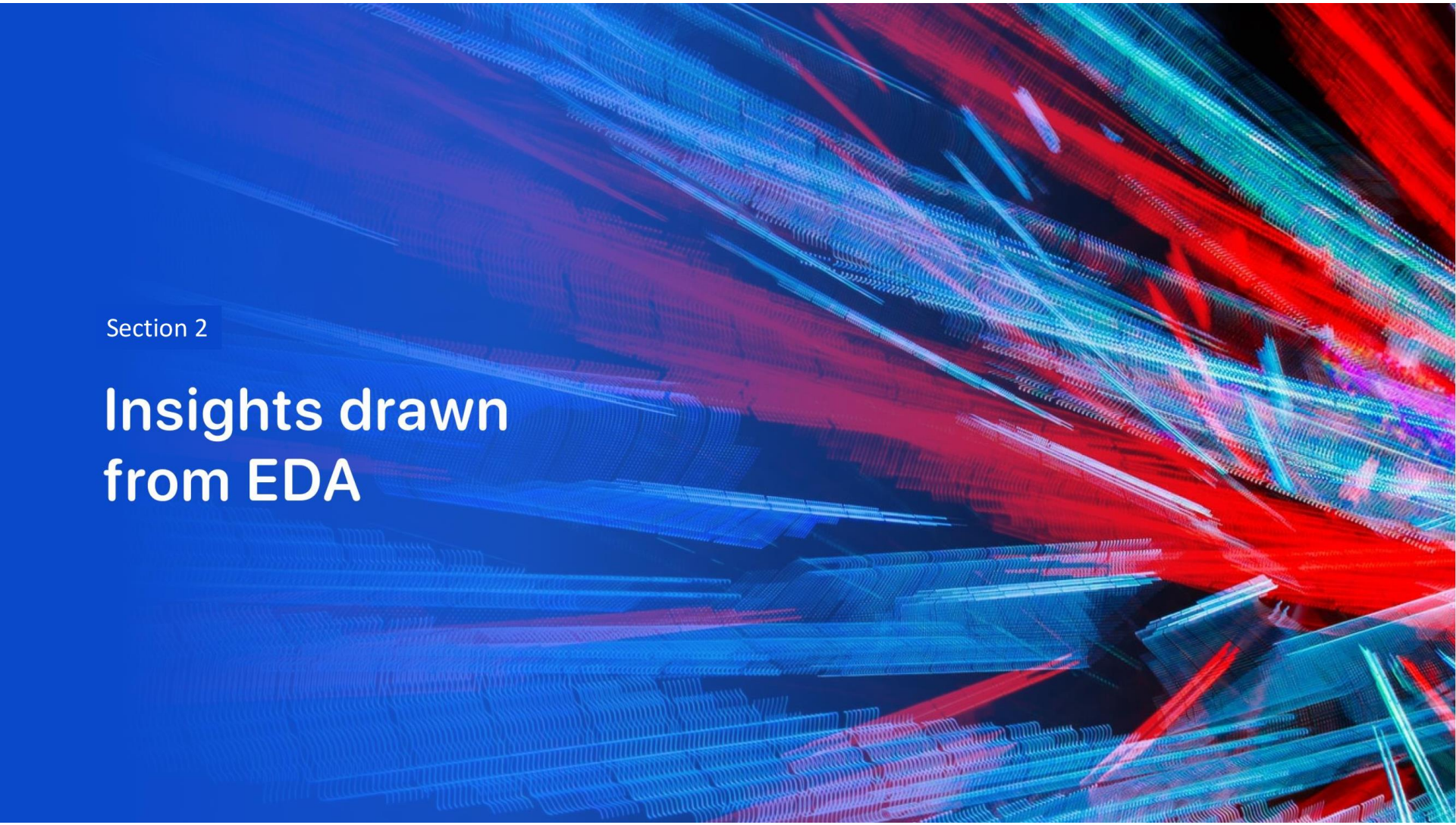
Success landing rates for different orbits

Success rate over the last 10 years

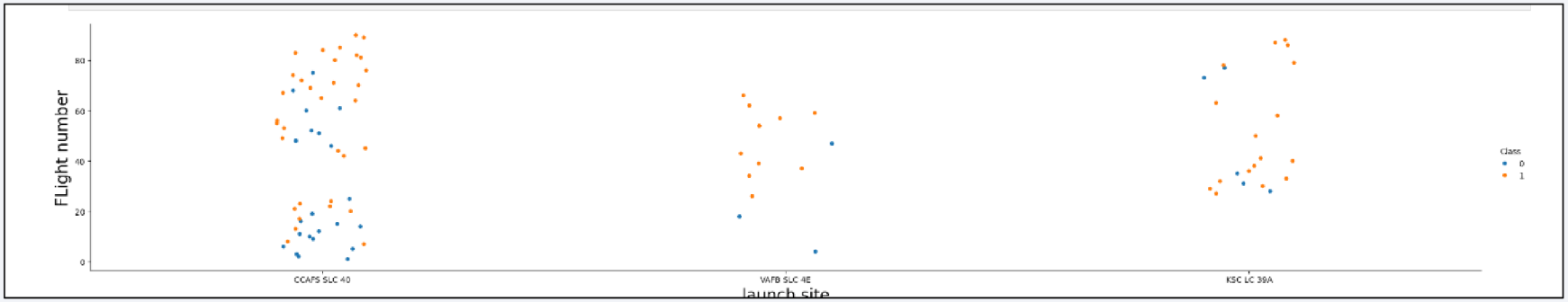




Plotly dash for different launch sites

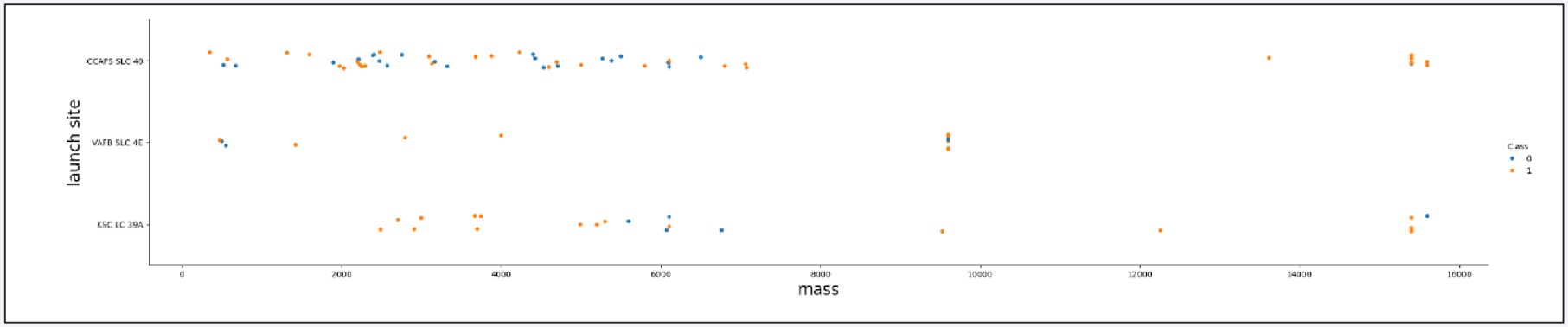

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- Launch site CCAFS had more launches, compared to VAFB and KSC. And a lower landing success rate.
- VAFB had the least number of unsuccessful landings.

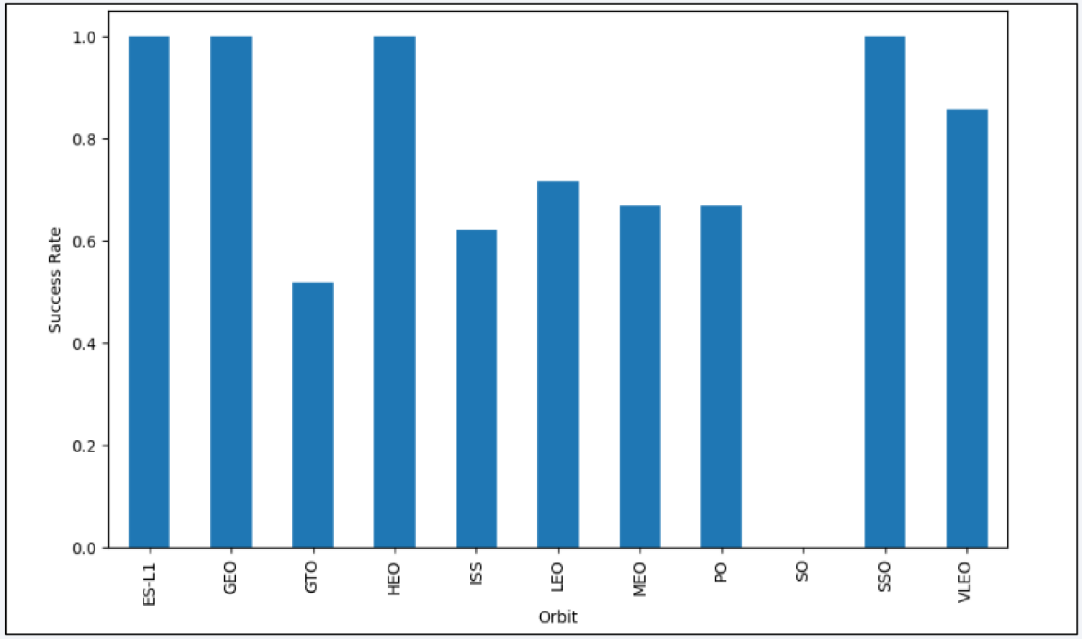
Payload vs. Launch Site



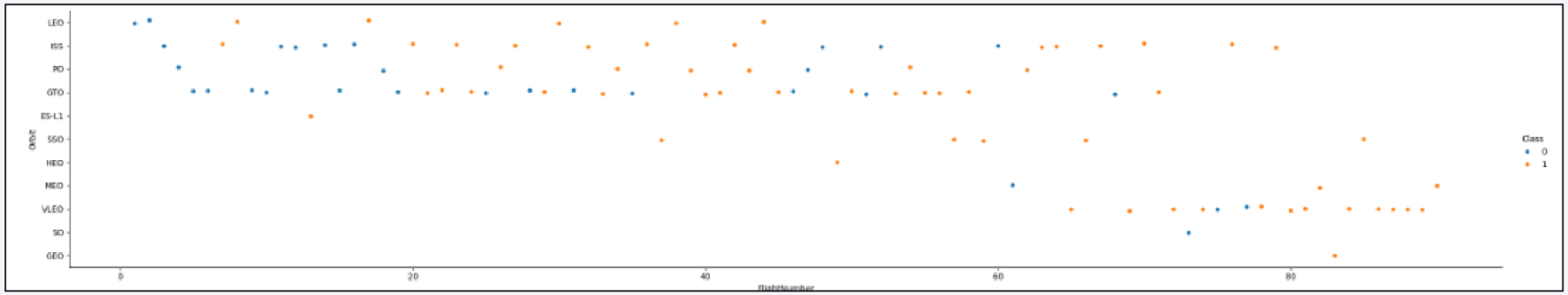
- VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000 kg).

Success Rate vs. Orbit Type

- SO orbit has 0 % success rate
- ES-L1,GEO,HEO and SSO have 100% success rates

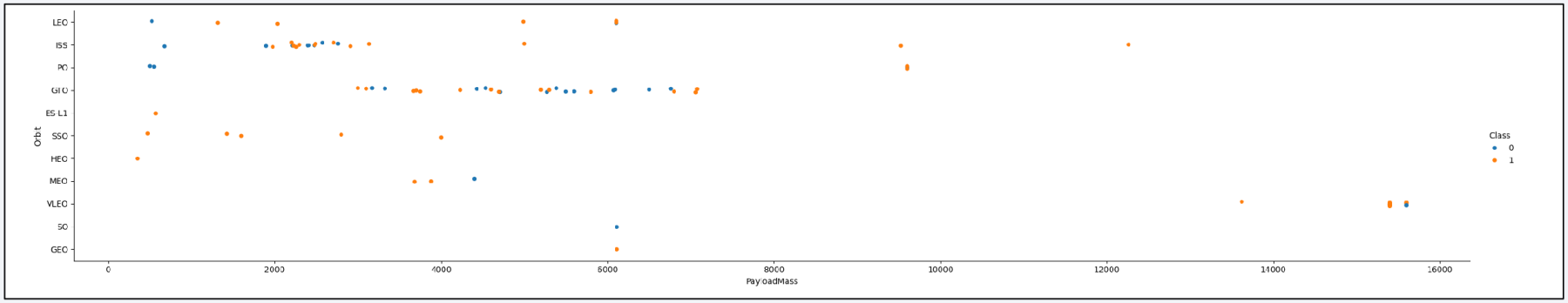


Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

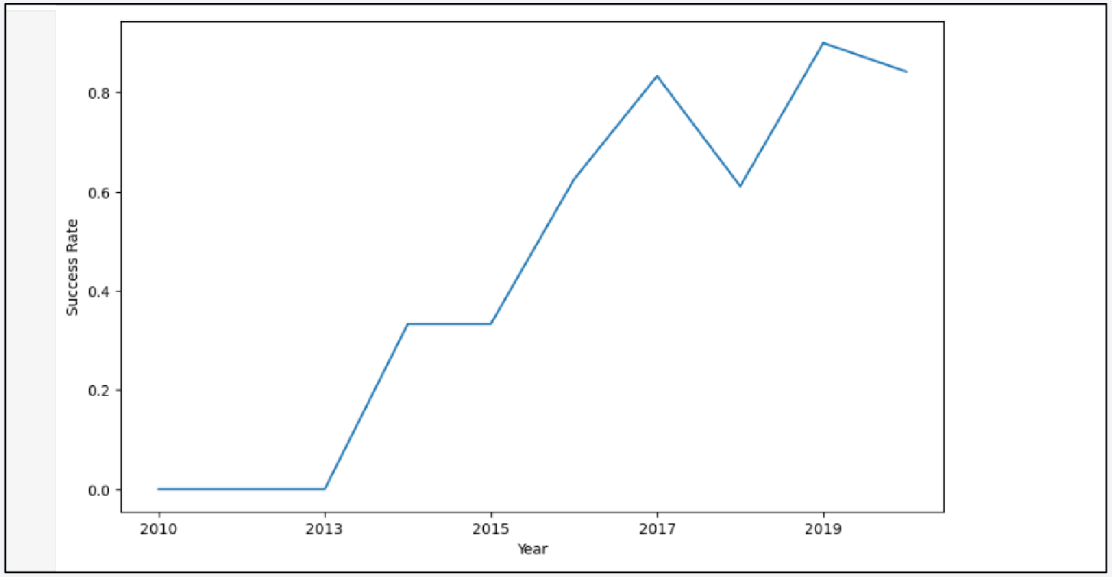
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Success rate increases on average over the 10 years.
- With the only significant decrease coming after 2017-2018



All Launch Site Names

```
*[66]: SELECT DISTINCT "Launch site" FROM my_data1_table;
```

[66]:

	Launch site
0	CCAFS
1	VAFB
2	Cape Canaveral
3	KSC
4	CCSFS
5	NaN

The unique launch sites listed in the result demonstrate the different number of locations used by SpaceX for launching missions. This helps in understanding the capabilities of SpaceX's launch infrastructure.

Launch Site Names Begin with 'CCA'

- The first 5 records from launch sites beginning with "CCA" show a mix of successful launches and "No attempt" outcomes for booster landings. This insight highlights the varied outcomes of missions from these specific launch sites.

[70]:
```
SELECT * FROM my_data1_table
WHERE "Launch site" LIKE 'CCA%'
LIMIT 5;
```

[70]:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	F9 v1.0B0003.1	Failure	4 June 2010	18:45
2	1	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0003.1	No attempt\n	4 June 2010	18:45
3	2	CCAFS	Dragon	525 kg	LEO	NASA	Success\n	F9 v1.0B0004.1	No attempt	8 December 2010	15:43
4	3	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44

Total Payload Mass

- The total payload mass carried by NASA boosters, amounting to 108,380 kg, highlights the contribution of NASA missions to space exploration efforts. This payload capacity demonstrates NASA's capability to handle a wide variety of missions.

```
*[76]: SELECT SUM(CAST(REPLACE(REPLACE("Payload mass", ' kg', ''), ',', '') AS FLOAT)) AS total_payload
       FROM my_data1_table
       WHERE "Customer" = 'NASA';
```

```
[76]: 108380.0
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by the booster version F9 v1.1 is approximately 2935.86 kg. This average indicates the typical payload capacity handled by this specific booster version.

```
SELECT AVG(CAST(REPLACE(REPLACE("Payload mass", ' kg', ''), ',', '') AS FLOAT)) AS average_payload
FROM my_data1_table
WHERE "Version Booster" = 'F9 v1.1';
average_payload
```

First Successful Ground Landing Date

- The earliest successful landing on a ground pad occurred on June 4, 2010. This milestone marks the beginning of successful spaceship recovery efforts, which are crucial for reusability and reducing the cost of space missions.

```
SELECT MIN(Date) AS first_successful_landing_date
FROM my_data1_table
WHERE "Launch outcome" LIKE '%Success%' AND "Launch outcome" LIKE '%ground pad%';
```

```
Date of the first successful landing on a ground pad: 2010-06-04 00:00:00
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass gre n 4000 but less than 6
- Present your query result wit ere

Total Number of Successful and Failure Mission Outcomes

- The overwhelming number of "Success" outcomes compared to "Failure" demonstrates SpaceX's high success rate in launching missions. This success rate is indicative of the reliability and efficiency of SpaceX's launch systems.

```
SELECT "Launch outcome", COUNT(*) AS count
FROM my_data1_table
GROUP BY "Launch outcome";
```

```
Success\n    72
Success      32
Failure       1
Name: Launch outcome, dtype: int64
```

Boosters Carried Maximum Payload

This query finds the maximum payload mass carried and then lists the names of the boosters that carried this maximum payload. It first determines the maximum payload mass using a subquery and then selects the Version Booster names for records that match this maximum payload mass.

The boosters listed in this result have achieved the remarkable feat of carrying the maximum payload mass, which was 15600 kg. This indicates these booster versions are among the most capable in terms of payload capacity, showcasing the advancements in SpaceX's booster technology.

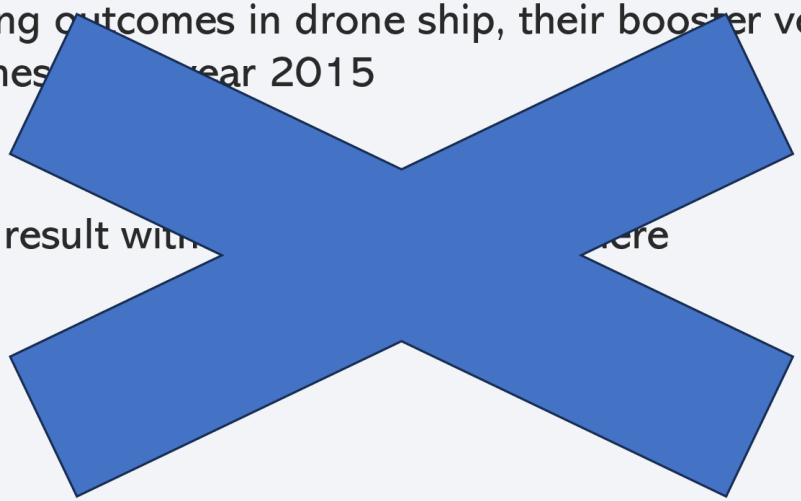
```
SELECT "Version Booster"
FROM my_data1_table, max_payload
WHERE CAST("PAYLOAD_MASS_KG" AS FLOAT) = max_payload_mass;
```

```
Maximum Payload Mass: 15600
Names of boosters which have carried the maximum payload mass:
['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

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

- Present your query result with a summary



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

```
SELECT "Landing_Outcome", COUNT(*) AS outcome_count
FROM my_data1_table
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY outcome_count DESC;
```

```
Ranked count of landing outcomes between 2010-06-04 and 2017-03-20:
No attempt               10
Failure (drone ship)      5
Success (drone ship)      5
Controlled (ocean)        3
Success (ground pad)      3
Failure (parachute)       2
Uncontrolled (ocean)      2
Precluded (drone ship)    1
```

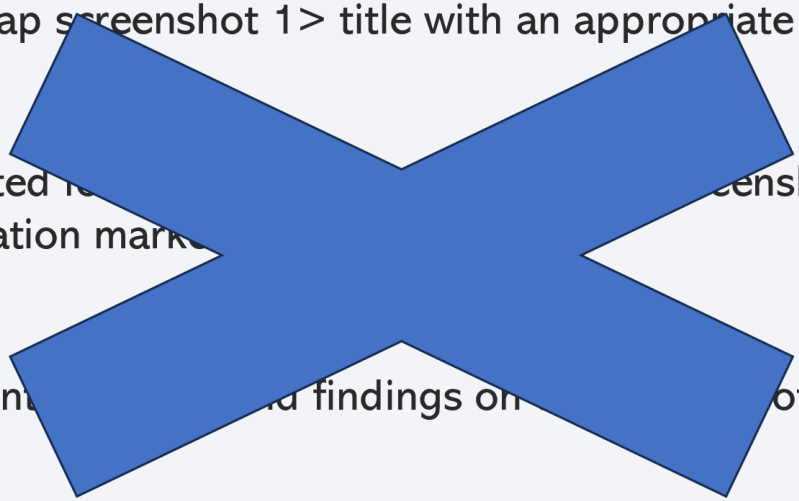
This query counts the occurrences of each unique landing outcome between June 4, 2010, and March 20, 2017. It groups the results by the Landing Outcome column and orders them in descending order by the count. The result ranks the landing outcomes based on their frequency within the specified date range.

Section 3

Launch Sites Proximities Analysis

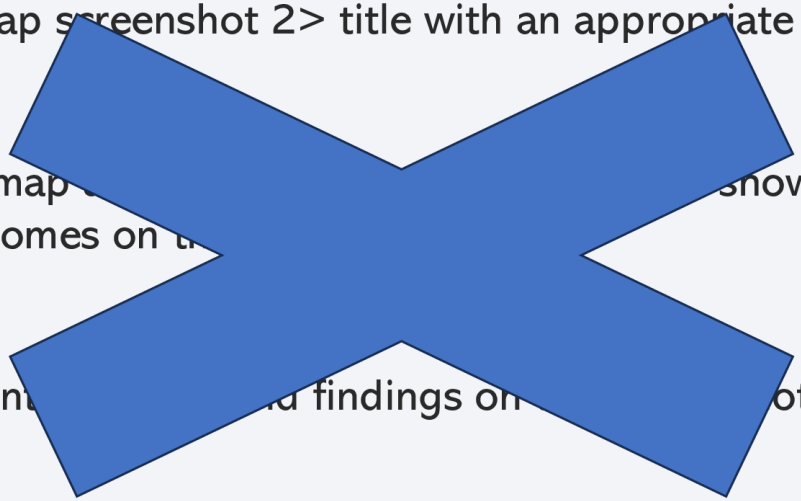
<Folium Map Screenshot 1>

- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated f… …enshot to include all launch sites' location mark…
- Explain the important… …d findings on… …ot



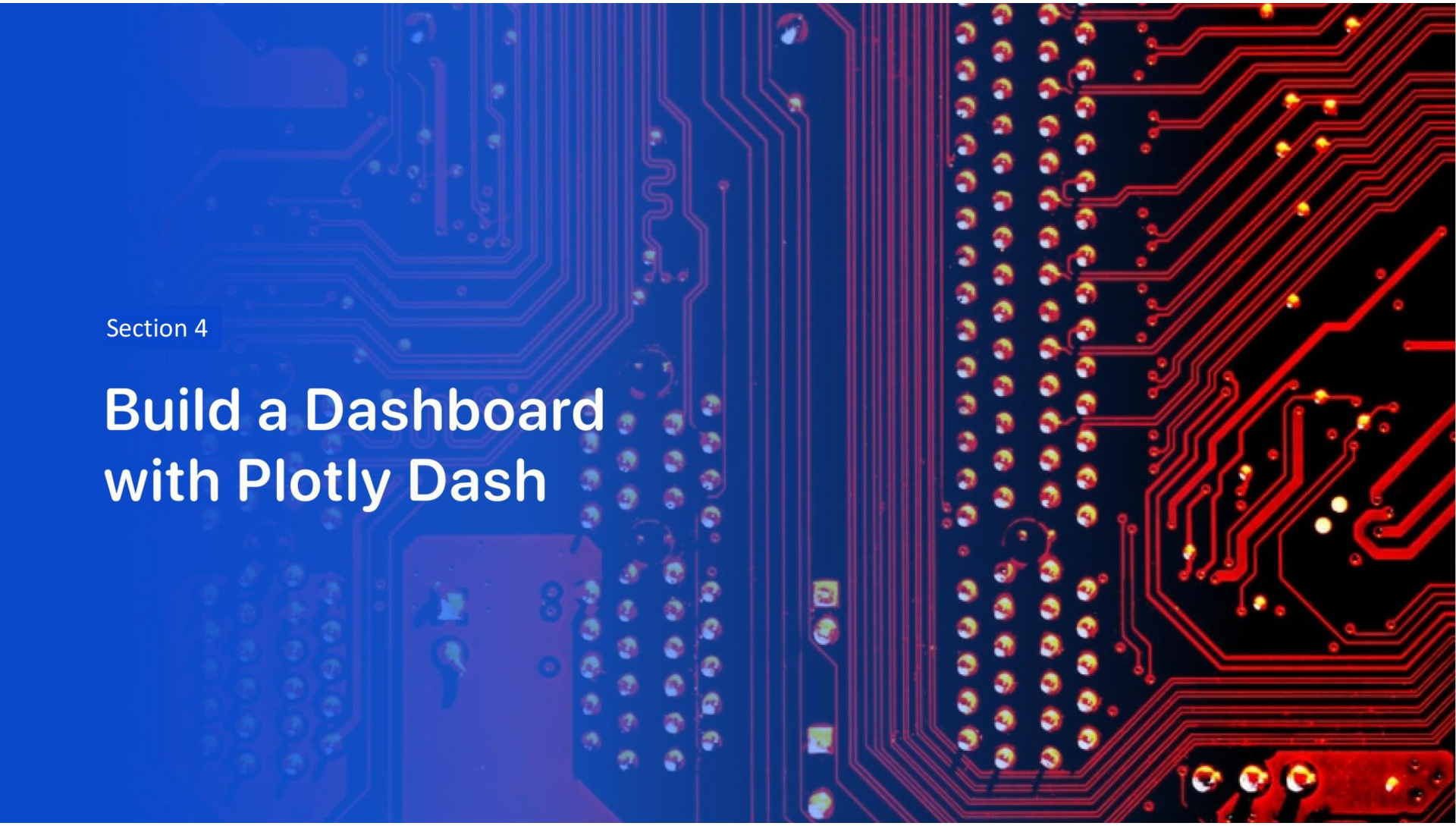
<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map … show the color-labeled launch outcomes on t…
- Explain the important… findings on … ot



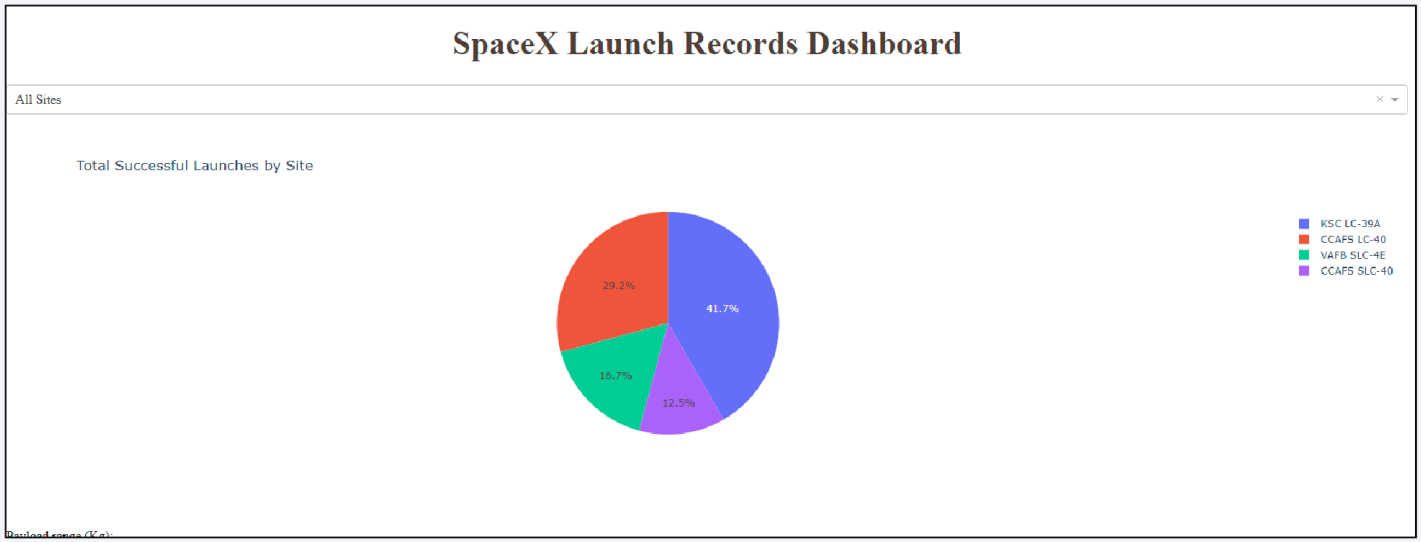
<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated … of a selected launch site to its … highway, coastline, with distance calcula…
- Explain the important … and findings on the … not


Section 4
Build a Dashboard with Plotly Dash

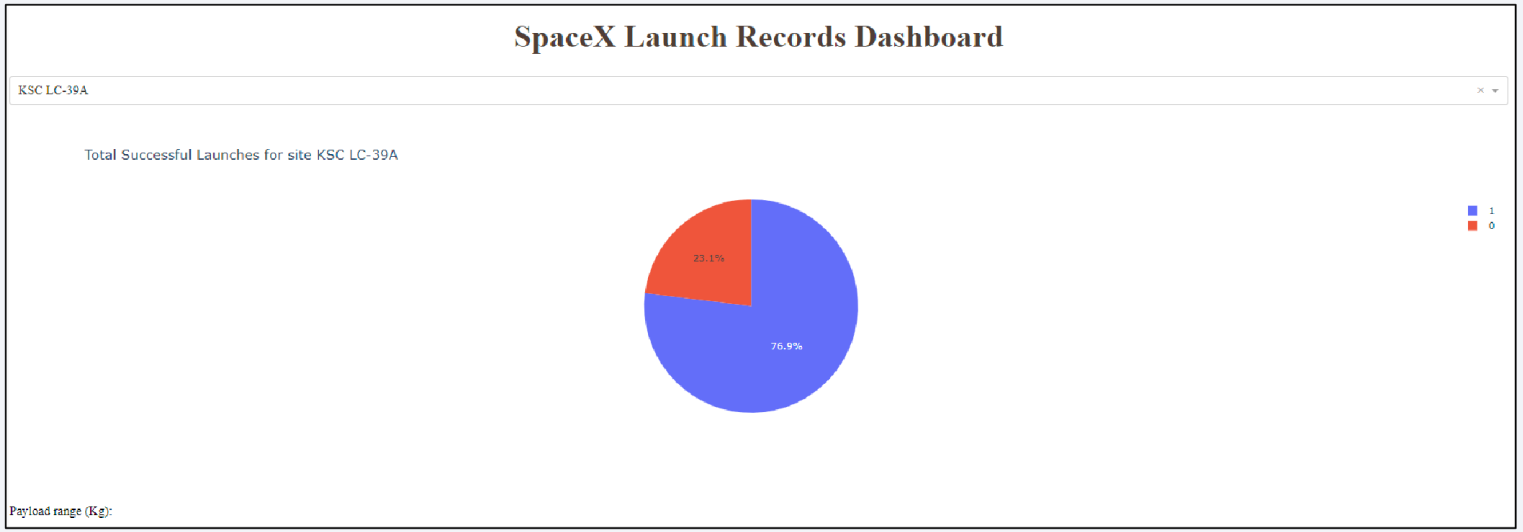
Launch success count for all sites

Launch site KSC LC-39A had the highest success rate out of all the sites



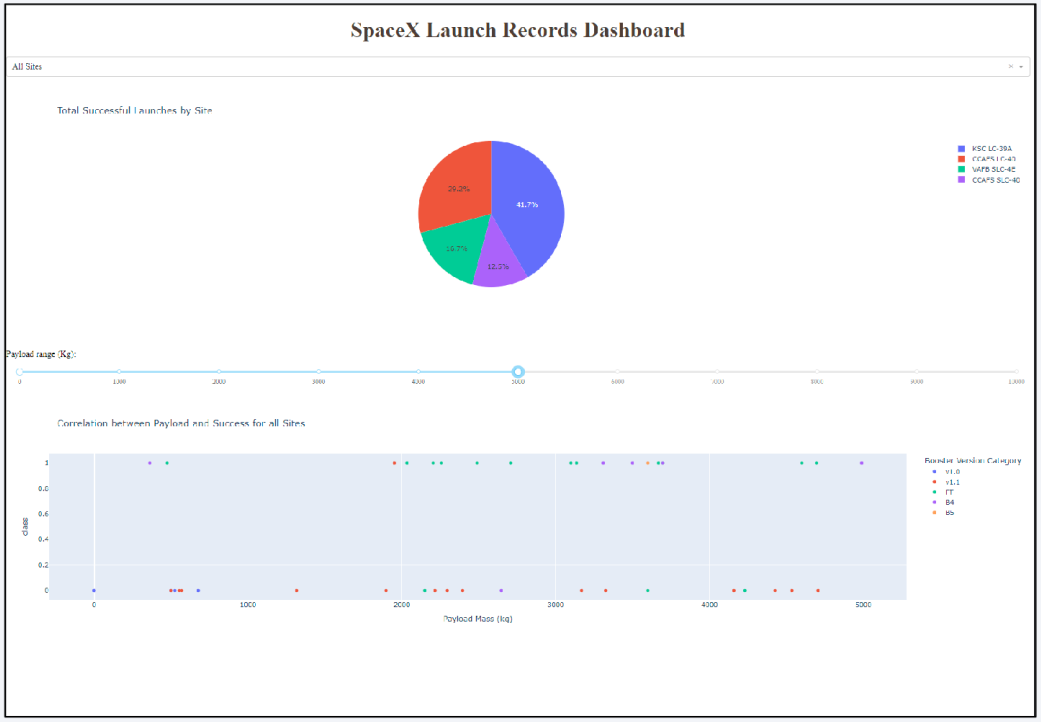
Site KSC LC-39A success ratio

- This shows the success rate for the most successful launch site (23.1% fail rate and 76.9% success rate)



Scatter plot for all sites vs Payload Mass

- Shows that for booster FT the success rate is high for payloads around 2500-3500 kg
- B4 boosters have a good success rate at payloads around 3500 kg

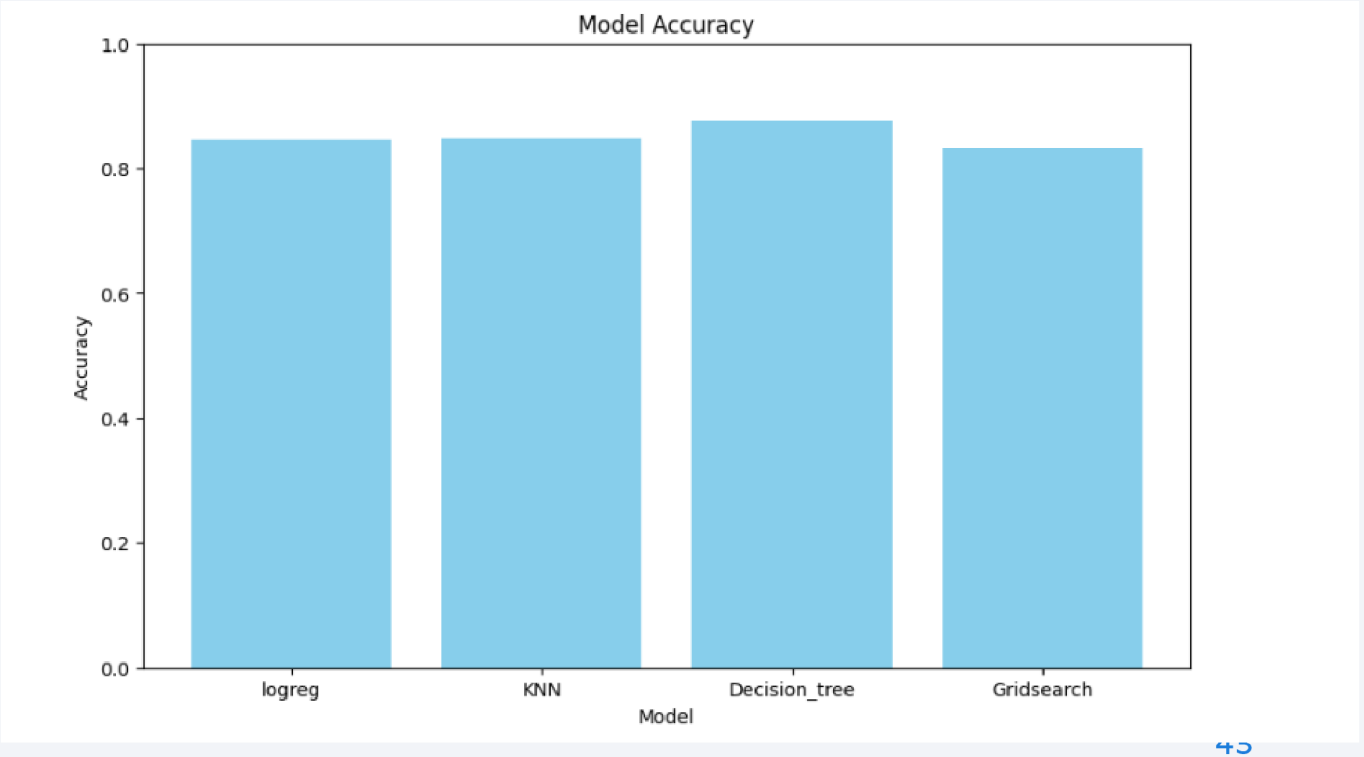


Section 5

Predictive Analysis (Classification)

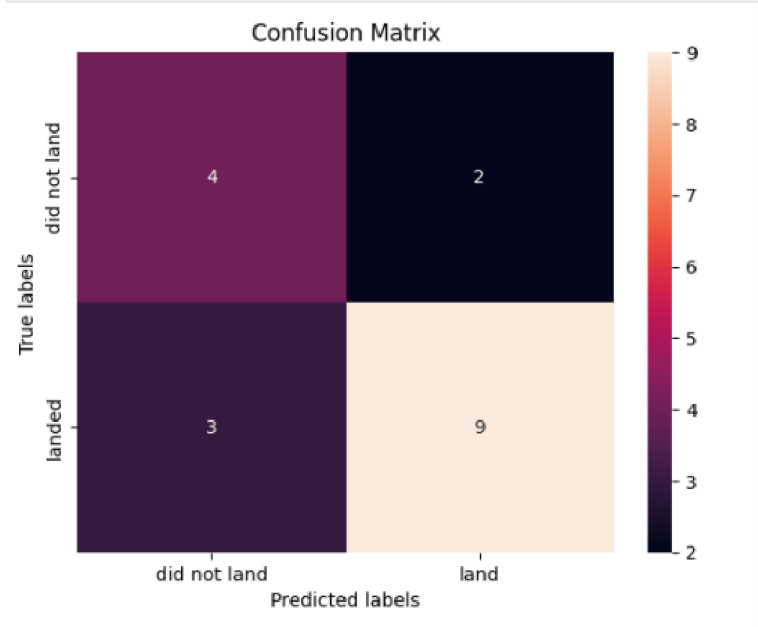
Classification Accuracy

- From the models built. Decision tree classifier had the greatest accuracy of 87%.



Confusion Matrix

- Decision tree had the best performing accuracy. And successfully predicted the most ‘Did not land’ events out of all the models.



Conclusions

- The analysis of model accuracies reveals that the Decision Tree Classifier had the highest accuracy out of all the models, achieving an accuracy of 87.5%. This performance can be attributed to its capability to better predict events where the rocket did not land successfully.
- In addition to model performance, an analysis of launch site success rates shows that the Kennedy Space Center (KSC LC-39A) had the most successful launches. This makes KSC the most reliable site for SpaceX launches in terms of success rate.
- Furthermore, when examining the success of launches by orbit type, it was found that the most successful orbits were Sun-Synchronous Orbit (SSO), Geostationary Orbit (GEO), Earth-Sun L1 (ES-L1), and Highly Elliptical Orbit (HEO). These orbits saw the highest number of successful launches, indicating a strong performance in missions targeting these specific orbits

Appendix

Thank you!