IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jochen Michalzik
13-February-2023

Outline

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

Executive Summary

Our objective in this capstone project is to predict the success or failure of the landing of SpaceX Falcon 9 first stage, which in turn will allow us to estimate the cost of a launch. To achieve this, we will use various machine learning classification techniques.

The project will be carried out in several phases including Data Collection, Data Cleaning and Preparation, Exploratory Data Analysis, Visualization, and finally, Machine Learning Modeling.

Our analysis will show that there are several factors related to rocket launches that have a strong influence on the success or failure of the landing. Ultimately, we will conclude that the Decision Tree algorithm may be the most suitable approach for solving this problem.

Introduction

The objective of this capstone project is to make accurate predictions about the landing of the Falcon 9 first stage by SpaceX. The company's ability to reuse the first stage has been a key factor in their ability to offer rocket launches at a fraction of the cost of their competitors.

By determining the likelihood of a successful landing, we can estimate the cost of a launch, which could be useful for companies considering bidding against SpaceX for a rocket launch opportunity.

Therefore, the central question that we aim to answer through this project is: Given a set of relevant features, will the first stage of a Falcon 9 rocket launch land successfully?

Section 1

Methodology

Methodology

Executive Summary

The data for this capstone project was obtained through two sources: the SpaceX API and a Wikipedia page. This data was then preprocessed using Python's pandas library to clean and transform the data.

Next, an Exploratory Data Analysis (EDA) was conducted using visualization tools like matplotlib, seaborn, and SQL queries to gain insights into the data. Interactive visualizations using Folium and Plotly Dash were also created to better understand the data.

Finally, we used four different machine learning classification algorithms - logistic regression, support vector machines, k-nearest neighbor, and decision tree classifier - to make predictions about the landing of the Falcon 9 first stage. These models were trained, optimized, and evaluated to determine the best performing one.

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- GitHub URL: Data Collection
- (https://github.com/Pansen123/Capstone_Course/blob/570bec807474d61f2c0ad6c7088401673ab2bdf4/01_jupyter-labs-spacex-data-collection-api_DONE.ipynb)

1. 1. Request and parse the SpaceX launch data using the GET
2. Normalize JSON response into a dataframe
3. Extract only useful columns using auxilary functions
4. Create new pandas dataframe from dictionary
5. Filter dataframe to only include Falcon 9 launches
6. Handle missing values
7. Export to CSV file

Data Collection - Scraping

- GitHub URL: Web Scraping
- https://github.com/Pansen123/Capstone_Course/blob/570bec807474d61f2c0ad6c7088401673ab2bdf4/02_jupyter-labs-webscraping_DONE.ipynb

1. Request rocket launch data from its Wikipedia page
2. Extract all column/variable names from the HTML table header
3. Create a data frame by parsing the launch HTML tables
4. Export to CSV file

Data Wrangling

- GitHub URL: Data Wrangling
- https://github.com/Pansen123/Capstone_Course/blob/570bec807474d61f2c0ad6c7088401673ab2bdf4/03_labs-jupyter-spacex-Data%2520wrangling_DONE.ipynb

1. **Calculate the number of launches on each site**
2. **Calculate the number and occurrence of each orbit**
3. **Calculate the number and occurence of mission outcome per orbit type**
4. **Create a landing outcome label from Outcome column using one-hot encoding**
5. **Export to CSV**

EDA with Data Visualization

- Different types of visualizations were used to represent and analyze the data:
- Scatter Plots: These plots were used to depict the relationship between two variables. For example, we compared the Flight Number with the Launch Site, Payload with the Launch Site, Flight Number with Orbit Type, and Payload with Orbit Type.
- Bar Charts: Bar charts were used to compare values across multiple categories. The x-axis represents the categories, and the y-axis represents a discrete value. For example, we used a bar chart to compare the success rate of different orbit types.
- Line Charts: Line charts are useful for demonstrating data trends over time. A line chart was used to show the success rate over a specific number of years.
- GitHub URL: EDA with data visualization

EDA with SQL

- Displaying the names of the distinct launch sites in the space mission
- Displaying 5 records where launch sites start with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying the average payload mass carried by the F9 v1.1 booster version
- Listing the date of the first successful landing outcome on a ground pad
- Listing the names of boosters that have had successful drone ship landings and have payload masses between 4000 and 6000
- Listing the total number of successful and failed mission outcomes
- Listing the names of the booster versions that have carried the maximum payload mass
- Listing the failed landing outcomes on drone ships, the associated booster versions, and launch site names for the year 2015
- Ranking the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.
- GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- Using the Folium library in Python, visualizations of launch sites and their success/failure outcomes were created on a map.
- Markers were used to represent each launch site and the success/failure of launches from that site, while lines were used to calculate the distances between each site and its surrounding locations.
- This analysis revealed interesting patterns about the locations of launch sites, such as their proximity to railways, highways, coastlines, and cities.
- GitHub: interactive map with Folium map

Build a Dashboard with Plotly Dash

- A Plotly Dash dashboard was created to display the results of the predictive analysis. This dashboard includes two main visualizations:
- A pie chart that displays the proportion of successful launches at each site. This chart provides a quick and easy way to understand the distribution of landing outcomes across all launch sites, or to see the success rate of launches at a specific site.
- A scatter plot that shows the relationship between the landing outcomes and the payload mass of various boosters. The dashboard takes two inputs: the site(s) and payload mass. This visualization allows users to see how different variables impact the landing outcomes.
- GitHub URL: plotly Dash lab

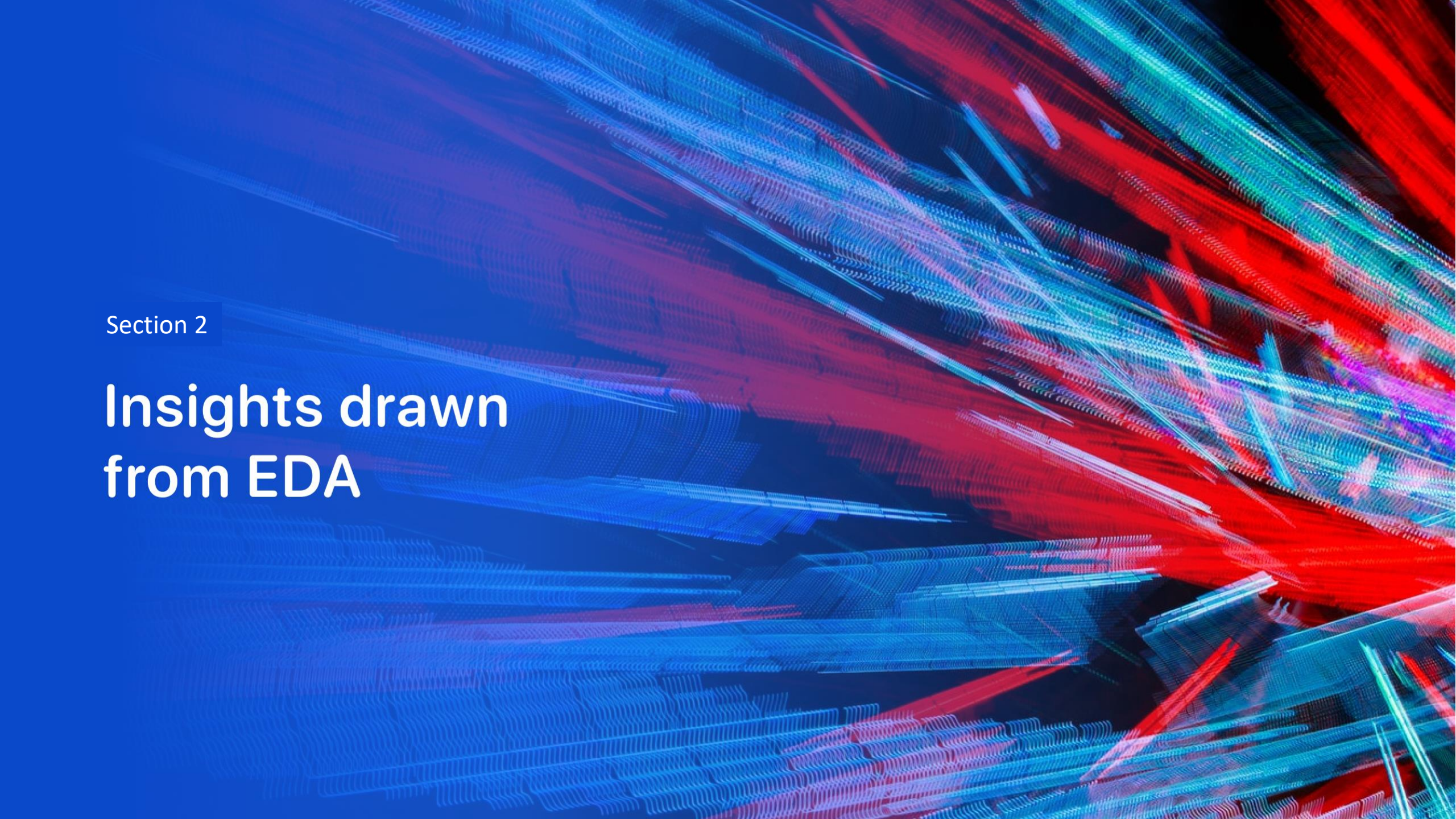
Predictive Analysis (Classification)

- GitHub URL: predictive analysis lab

1. Create column for “Class”
2. Standardizing the data
3. Split ito training and test set
4. Find best Hyperparameter for SVM, Decision Trees, K-Nearest Neighbours and Logistic Regression.
5. Use test data to evaluate models based on their accuracy scores and confusion matrix

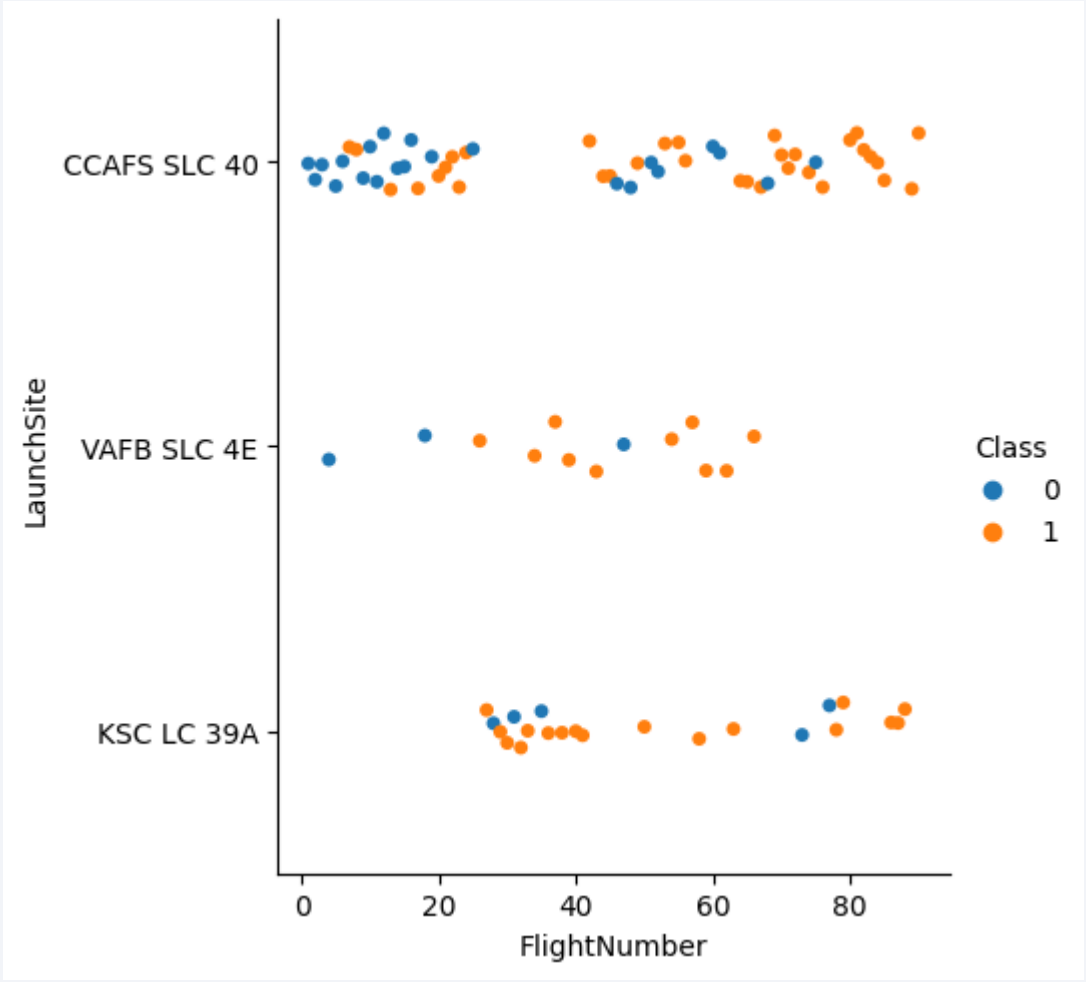
Results

- The results of the data analysis showed that the success rate of the Falcon 9 landings was found to be **66.66%.**
- The predictive analysis also yielded that the Decision Tree algorithm was the best classification method with a remarkable accuracy of **94%.**


Section 2
Insights drawn from EDA

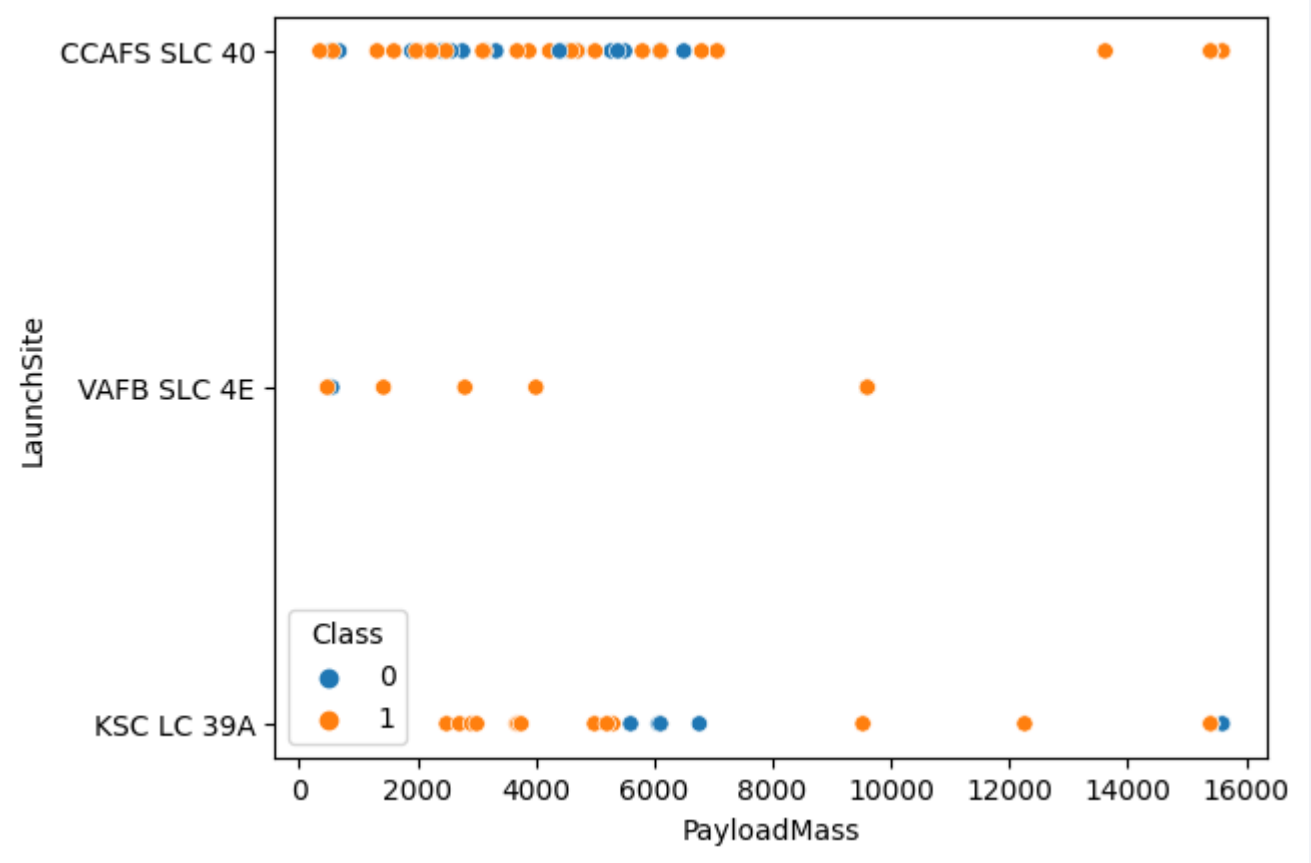
Flight Number vs. Launch Site

- This graph displays the relationship between the number of flights and the success rate.
- The successful launches are represented by blue dots while unsuccessful launches are marked with red dots.
- It is evident that there was a steady increase in the success rate as the number of flights increased, particularly after the 40th launch.



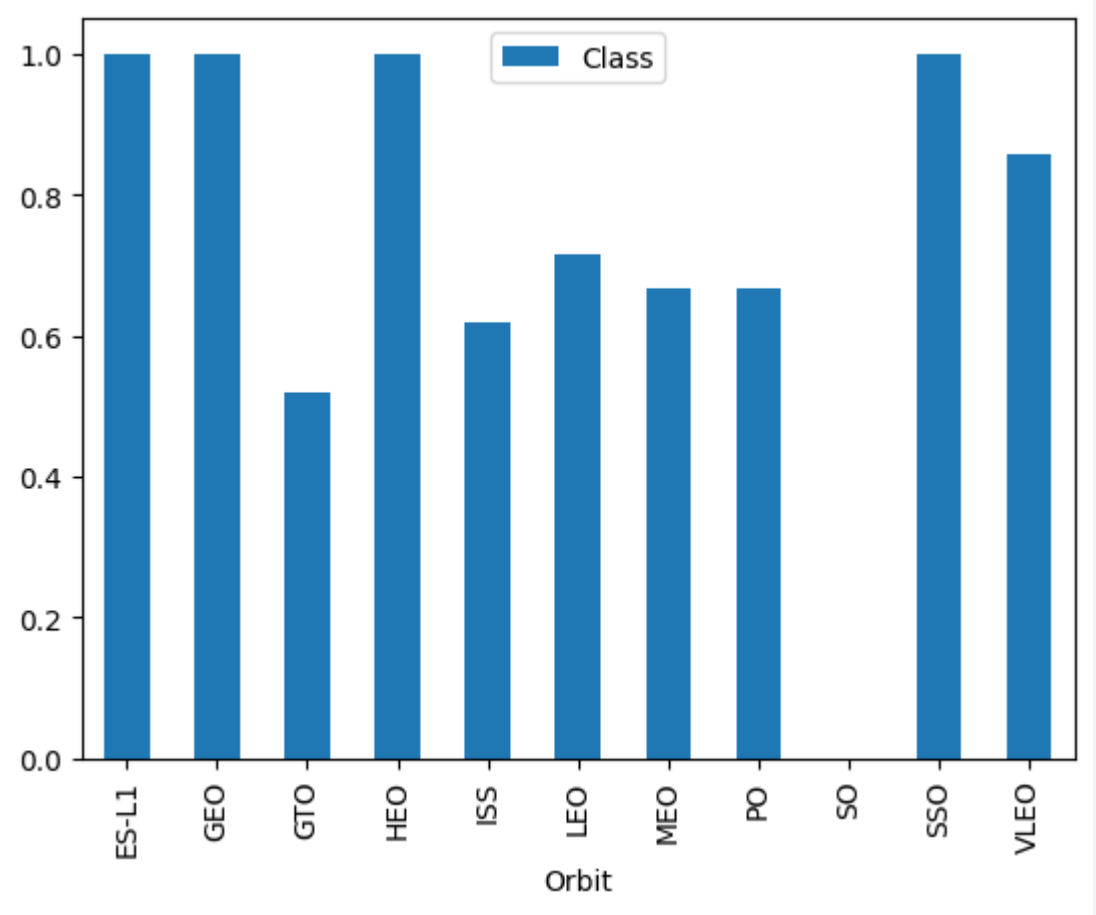
Payload vs. Launch Site

- The chart displays the relationship between launch outcomes and payload mass for a certain launch site.
- Blue dots signify successful launches, while red dots indicate failed missions.
- Notably, the chart suggests that there were no launches with heavy payloads from the VAFB-SLC site.
- Additionally, the correlation between payload mass and launch site appears to be weak, indicating that this metric may not be the best factor for making decisions.



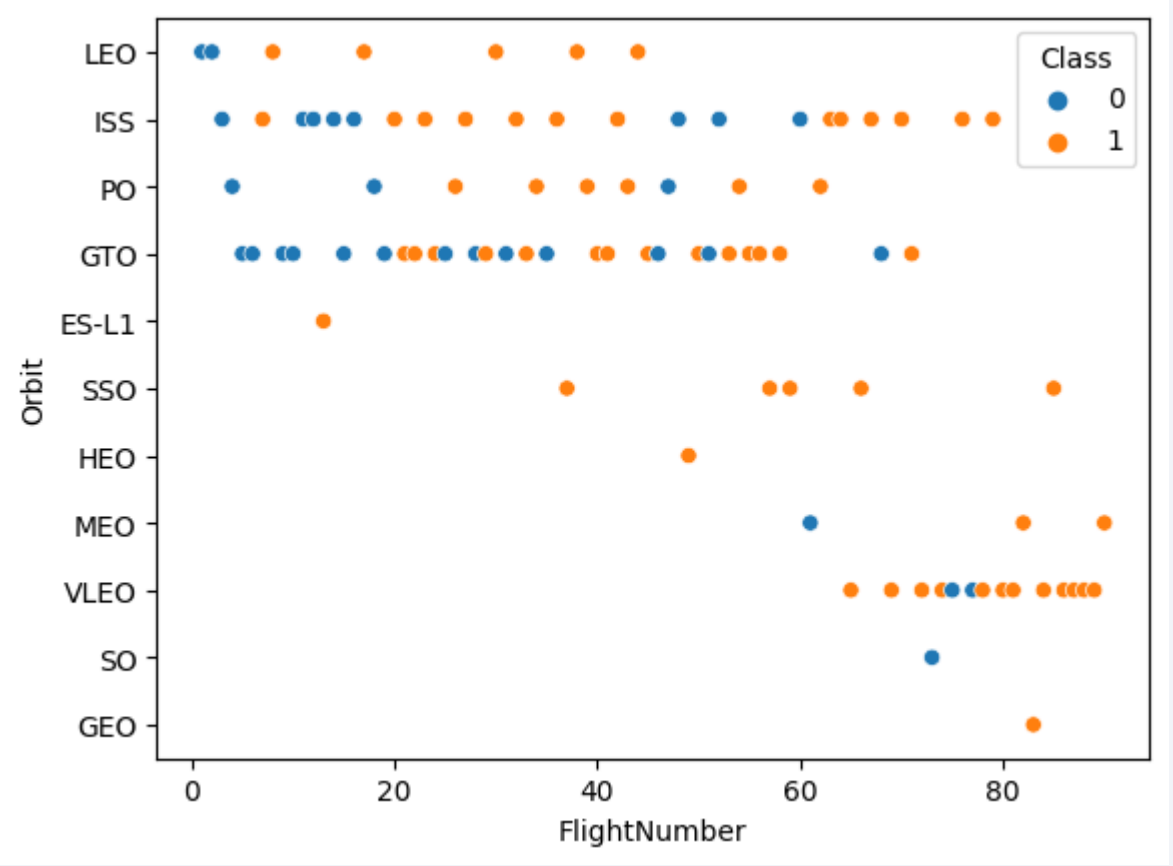
Success Rate vs. Orbit Type

- The success rates of the SSO, HEO, GEO, and ES-L1 orbits are all 100%.
- Conversely, the SO orbit has not seen any successful launches, resulting in a 0% success rate.



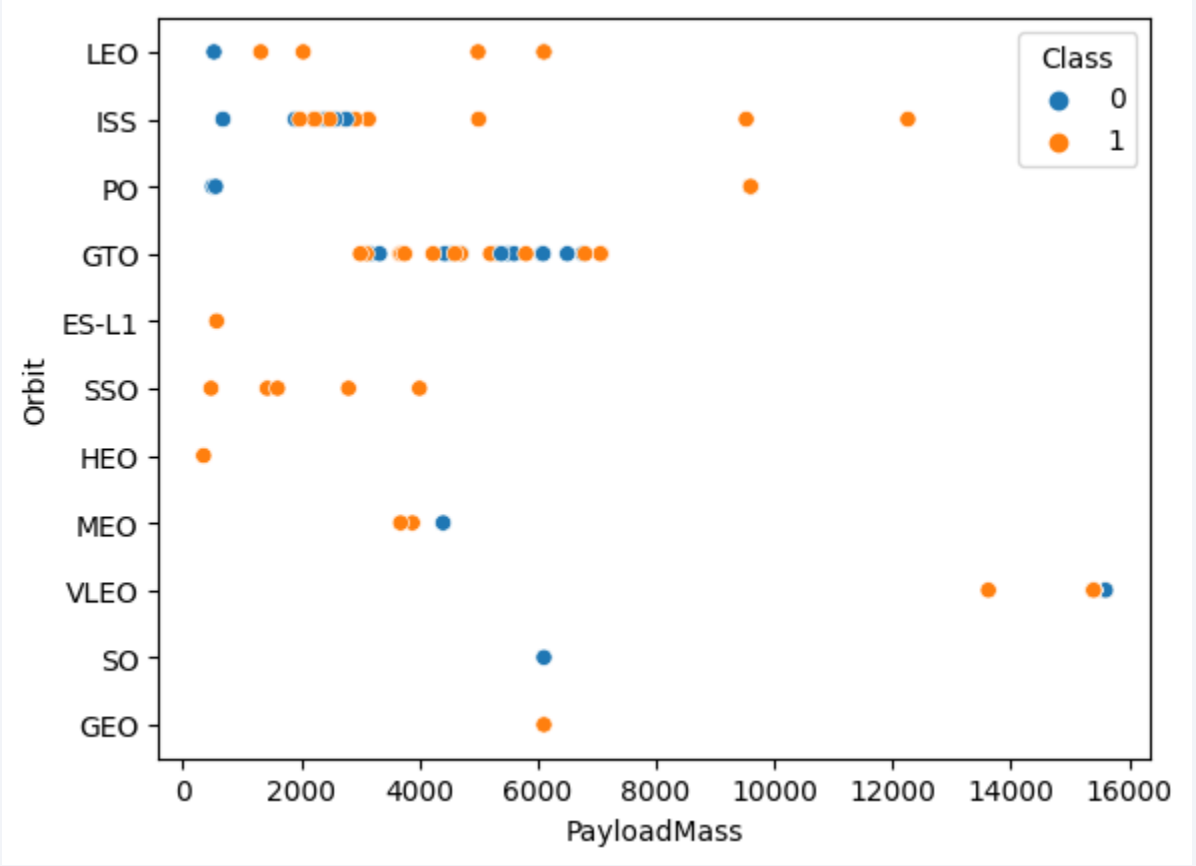
Flight Number vs. Orbit Type

- The correlation between the number of flights and success rate in the LEO orbit is positive. This means as the number of flights increases, the success rate also increases.
- However, the relationship between flight number and success rate in the GTO orbit is unclear and no clear pattern can be seen.
- Despite having a 100% success rate in the SSO orbit, there have been fewer flights compared to other orbits.
- Additionally, it can be seen that flights with flight numbers greater than 40 have a higher success rate compared to flights with flight numbers between 0 and 40.



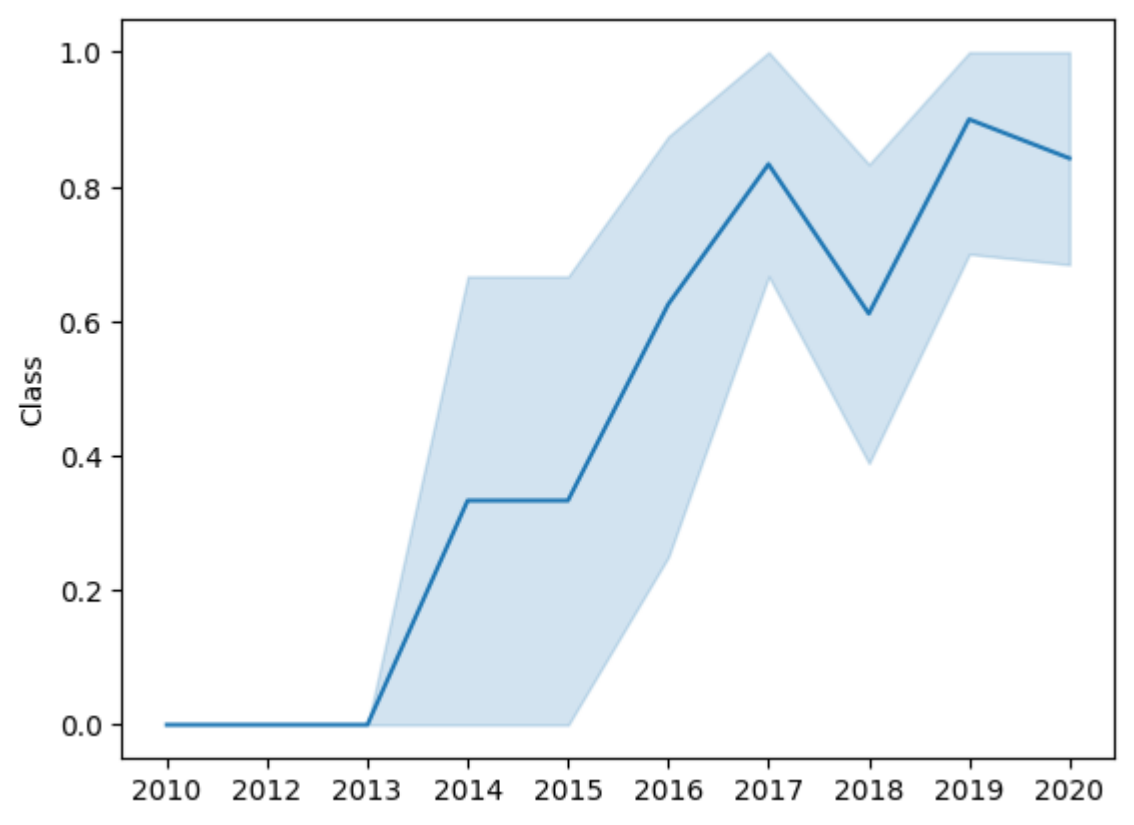
Payload vs. Orbit Type

- As the payload mass increases, the success rate in the PO, SSO, LEO, and ISS orbits also increases.
- There appears to be no clear correlation between orbit type and payload mass in the GTO orbit, as both successful and unsuccessful launches are present in equal measure.



Launch Success Yearly Trend

- The chart indicates a general upward trend in the success rate of landings over time. However, there are noticeable dips in success rate in the years 2018 and 2020.



All Launch Site Names

- The DISTINCT clause was used to return only the unique rows from the launch_site column.
- The names of the launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E .

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- The top five results of launch sites beginning with the string 'CCA' were displayed using the LIMIT and LIKE clauses.

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload carried by boosters from NASA was determined using the SUM() function, which was applied to the payload_mass__kg column.

	SUM(PAYLOAD_MASS__KG_)
0	107010

Average Payload Mass by F9 v1.1

- The AVG() function was employed to determine the average payload mass carried by boosters with the version F9 v1.1.
- The WHERE clause was utilized to limit the calculation to only those booster versions labeled as "F9 v1.1".

	AVG(PAYLOAD_MASS__KG_)
0	2534.666667

First Successful Ground Landing Date

- The MIN() function was utilized to determine the earliest date of a successful landing outcome on the ground pad.
- The WHERE clause was utilized to limit the results to only those entries where the value in the 'landing_outcome' column was specified as 'Success (ground pad)'.

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The WHERE clause was utilized to limit the results to only successful drone ship landings, while the BETWEEN clause was utilized to display only those with a payload mass in the range between 4000 and 6000 kilograms.

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

Total Number of Successful and Failure Mission Outcomes

- The GROUPBY clause was utilized with the COUNT() function to determine the frequency of different mission outcomes.
- The 'mission_outcome' column was the focus of the analysis, and the results showed the total number of successful and failed mission outcomes.
- Out of 101 missions, 99 were successful.

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The subquery employed the MAX() function to identify the boosters with the highest payload mass from the 'payload_mass__kg' column. A list of these boosters was then generated.

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

2015 Launch Records

- The SELECT statement was used to retrieve multiple columns from the table.
- The YEAR(DATE) function was used to retrieve only those rows with a 2015 launch date.

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The COUNT() function was utilized to determine the frequency of various landing outcomes. The BETWEEN and WHERE clauses were utilized to restrict the results to those between June 4, 2010 and March 20, 2017.
- The results were grouped according to their outcome through the use of the GROUPBY clause and finally, the ORDERBY and DESC clauses were utilized to sort the results in a descending order.

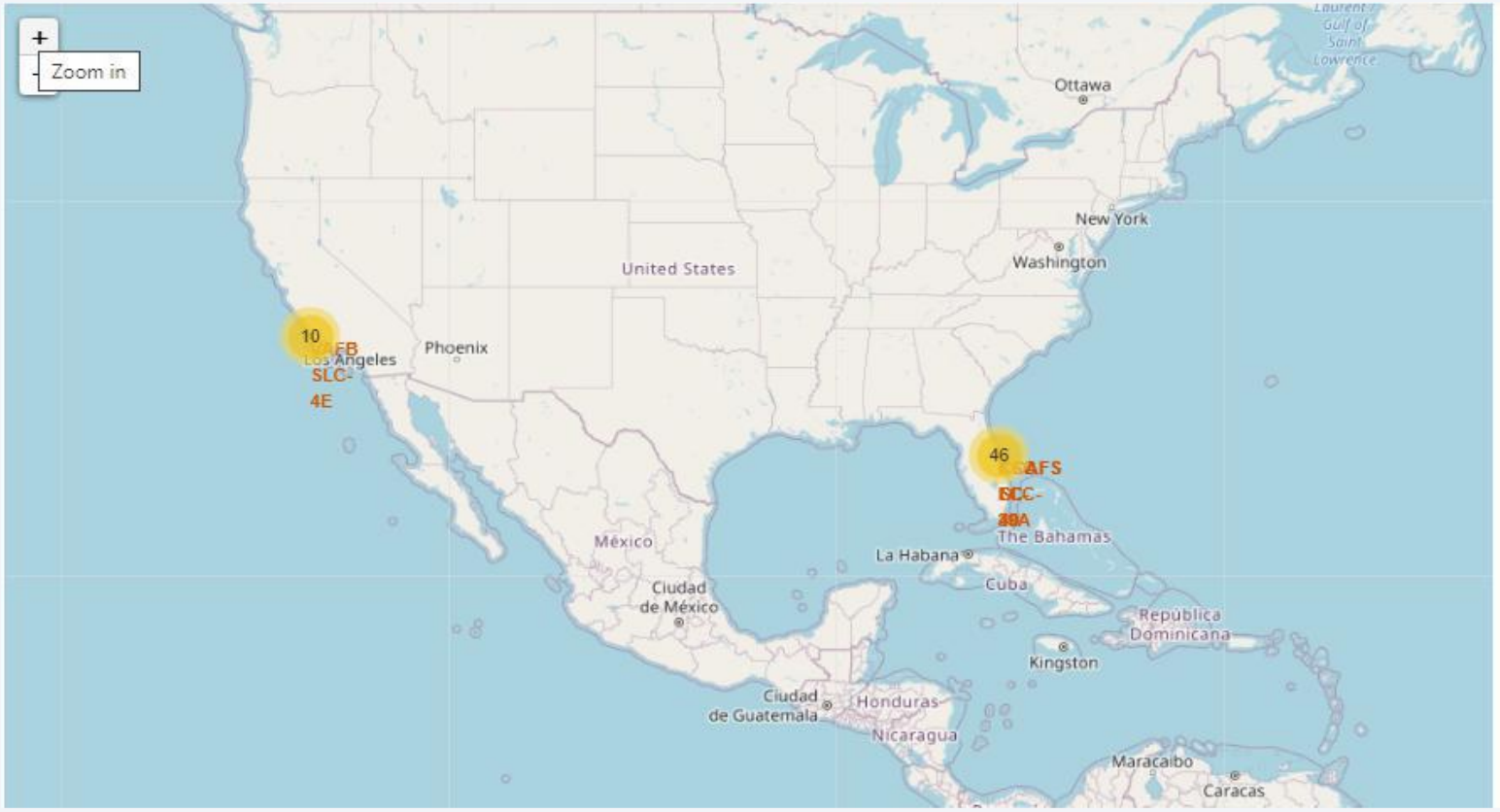
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

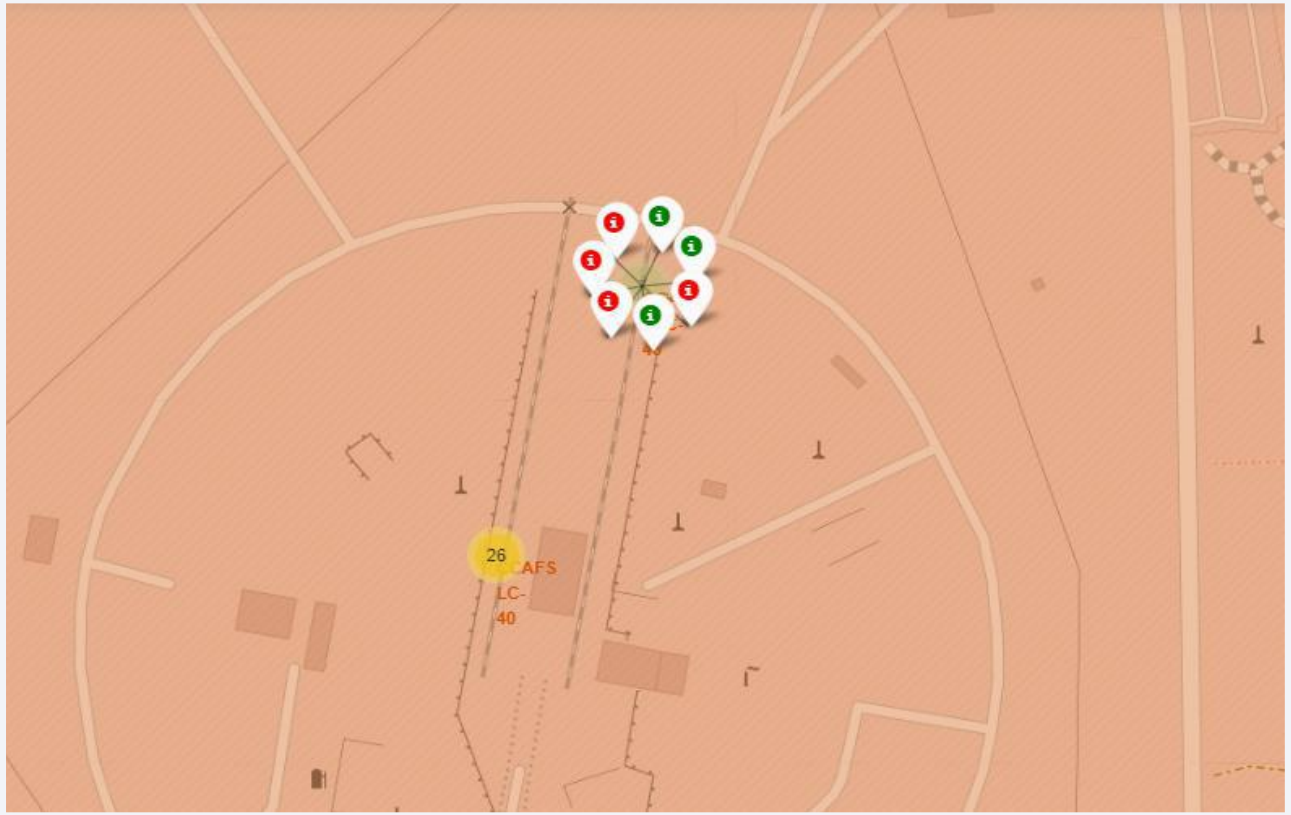
SpaceX launch sites

- The yellow markers on the map indicate the locations of all SpaceX launch sites in the United States.
- The launch sites have been positioned near the coast with a strategic approach.



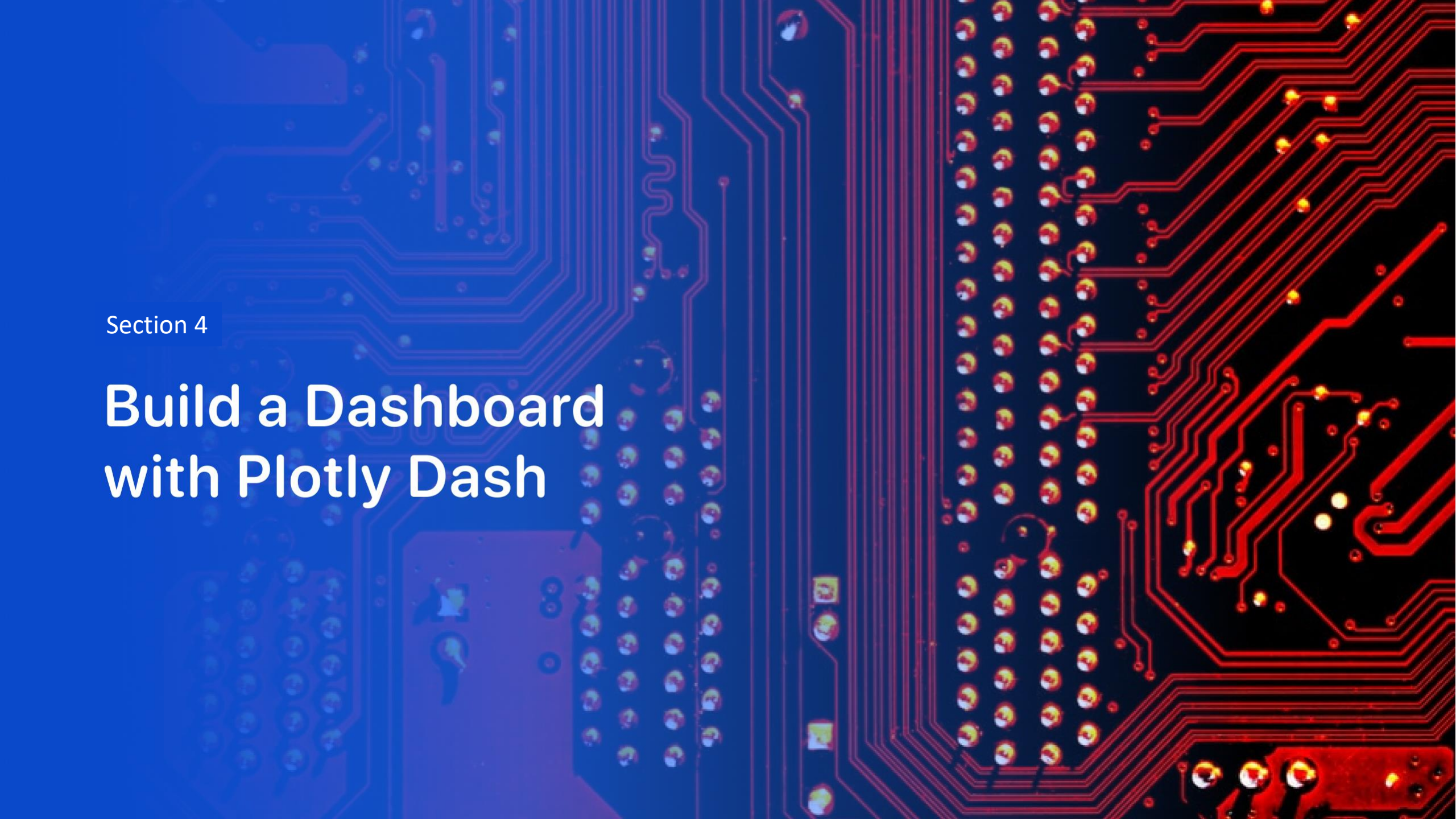
successful or failed landings

- When we zoom in on a SpaceX launch site, we have the ability to click on the site and view marker clusters which indicate successful landings (represented by green markers) or failed landings (represented by red markers).



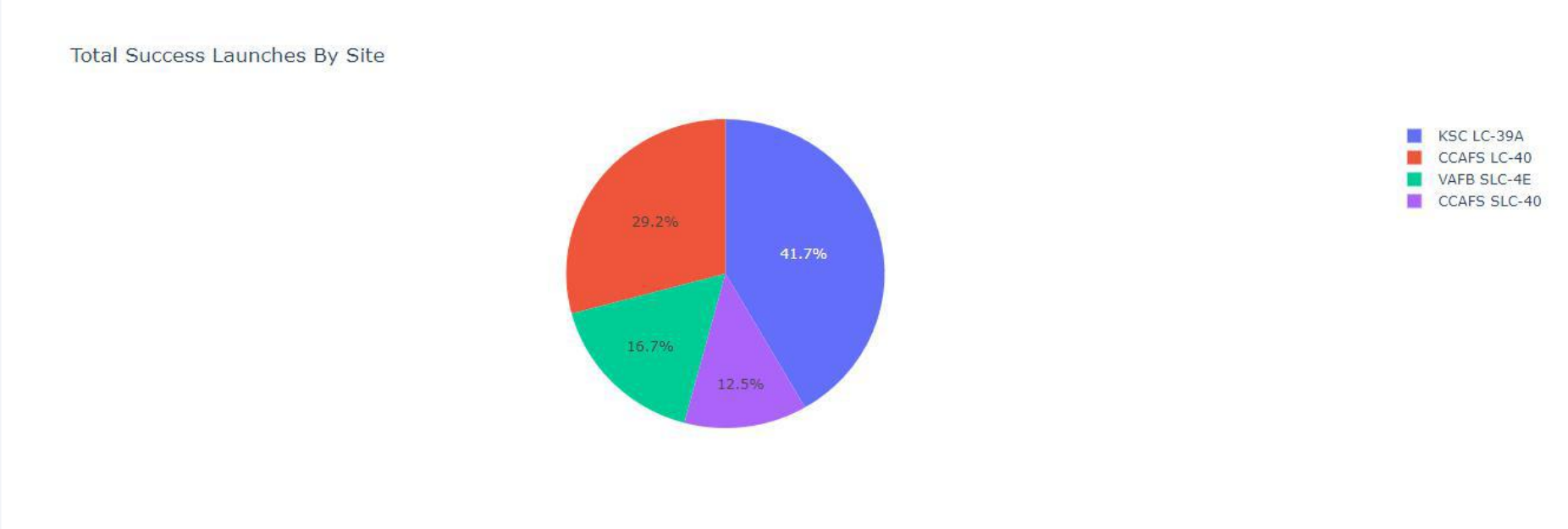
chosen launch site

- The map generated shows the proximity of the chosen launch site to key locations. The site is situated near a highway for easy transportation of personnel and equipment. Additionally, the site is situated near the coastlines to facilitate launch failure testing.
- To ensure safety, the launch sites are also kept at a sufficient distance from urban areas. This information can be viewed in the accompanying notebook.


Section 4
Build a Dashboard with Plotly Dash

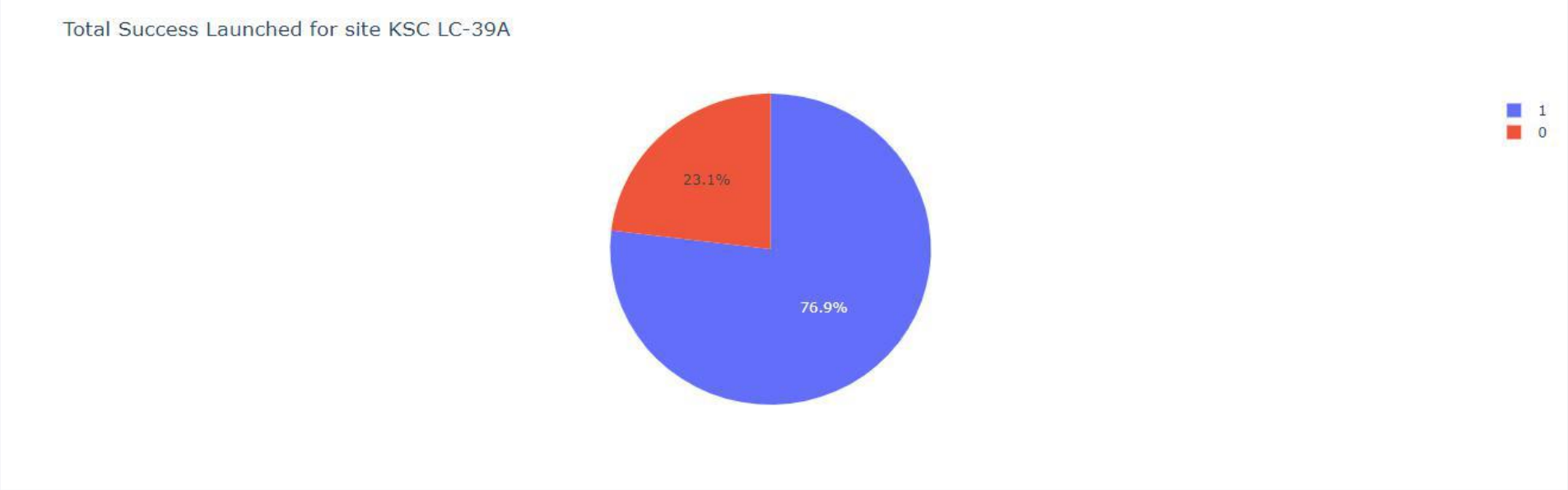
Launch Site at Kennedy Space Center

- The LC-39A Launch Site at Kennedy Space Center has recorded the highest number of successful launches, with a total of 10.



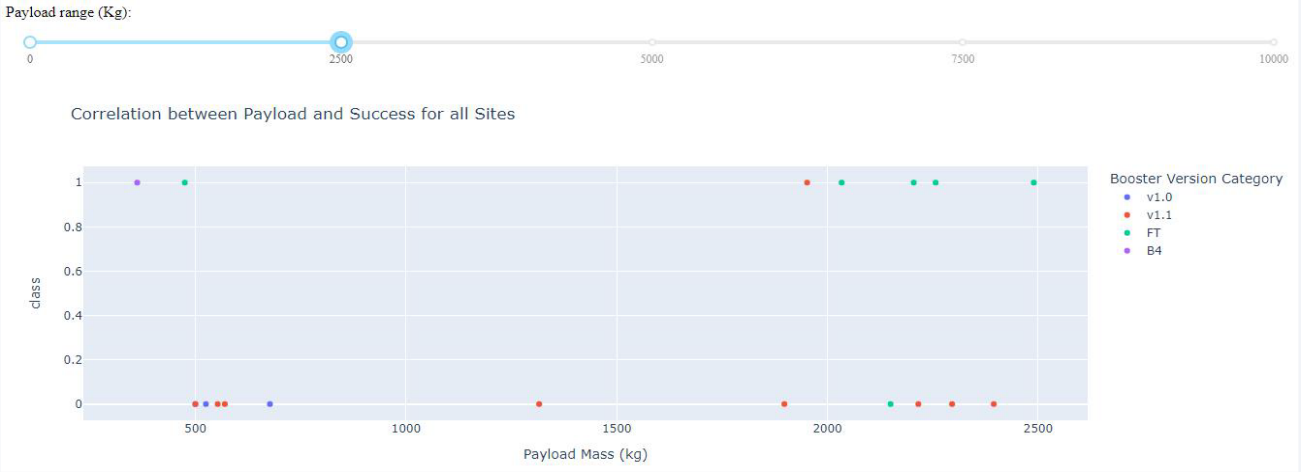
highest success rate

- The KSLC-39A has the highest success rate with 76.9%.



Payloads versus Launch Outcome

- The launch success rate for payloads 0-2500 kg is slightly lower than that of payloads 2500-5000 kg. There is in fact not much difference between the two.
- The booster version that has the largest success rate, in both weight ranges is the v1.1.



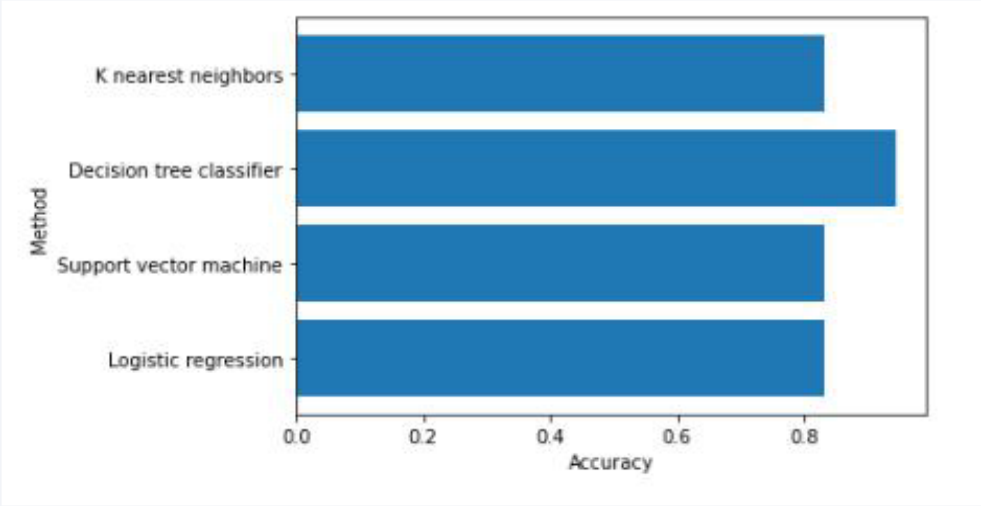
Section 5

Predictive Analysis (Classification)

Classification Accuracy

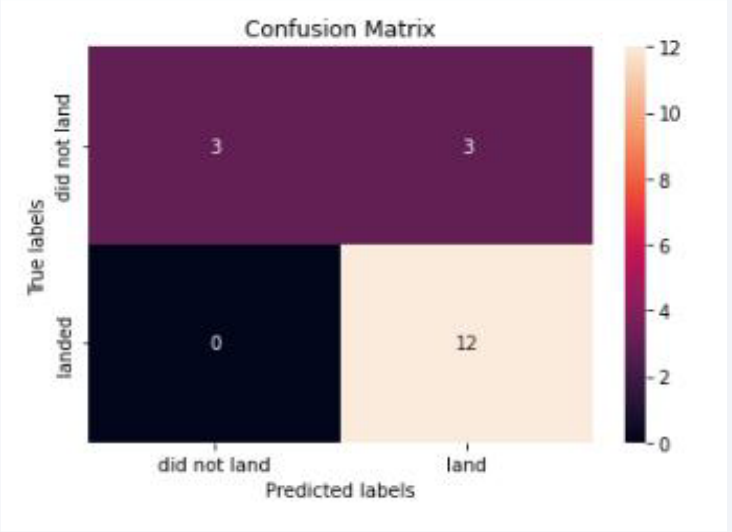
- The Decision Tree classifier had the best accuracy at 94%.

	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333



Confusion Matrix

- The model predicted 12 successful landings when the True label was successful (True Positive) and 3 unsuccessful landings when the True label was failure (True Negative).
- The model also predicted 3 successful landings when the True label was unsuccessful landing (False Positive).
- The model generally predicted successful landings.



Conclusions

- The analysis showed that there is a positive correlation between number of flights and success rate as the success rate has improved over the years.
- There are certain orbits like SSO, HEO, GEO, and ES-L1 where launches were the most successful.
- Success rate can be linked to payload mass as the lighter payloads generally proved to be more successful than the heavier payloads.
- The launch sites are strategically located near highways and railways for transportation of personel and cargo, but also far away from cities for safety.
- The best predictive model to use for this dataset is the Decision Tree Classifier as it had the highest accuracy with 94%.

Appendix

- Coursera Project Link: https://www.coursera.org/learn/applied-data-sciencecapstone/home/welcome

- GitHub Repository: https://github.com/Pansen123/Capstone_Course

Thank you!