IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ganesh Gangaram Shelar
10-08-2023

Outline

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

Executive Summary

- The project aimed to predict the success of **SpaceX rocket landings**. Data was collected from the public SpaceX API and SpaceX Wikipedia page. A 'class' column was created to classify successful landings.
- Data exploration was conducted using SQL, visualization, Folium maps, and dashboards. Relevant columns were selected as features. Categorical variables were converted to binary using one-hot encoding. Data was standardized for consistency.
- Four machine learning models were implemented:

    1. Logistic Regression

    2.Support Vector Machine

    3.Decision Tree Classifier

    4.K Nearest Neighbors

  All models achieved an accuracy rate of approximately **83.33%**.

- The project faced limitations due to data availability.Challenges included addressing data quality issues and achieving a balanced dataset.

Introduction

Welcome to the culmination of our IBM Data Science Capstone project—predicting the outcome of SpaceX rocket landings. In a world where space exploration pushes boundaries, our project leverages data science to forecast the success of these missions.

SpaceX's pursuit of innovation is inspiring, yet every launch carries inherent risks. Through meticulous data collection, feature engineering, and machine learning, we've developed models with an impressive accuracy rate of approximately 83.33%. However, as we'll discuss, our models have a tendency to over-predict successful landings, indicating a need for more data and refinement.

Join me as we explore the intersection of data science and space exploration.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - This data wrangling methodology ensured that the dataset was prepared systematically and efficiently for the predictive modeling phase of the project.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, evaluate classification models

Data Collection

In our project, we meticulously gathered data from two primary sources: the SpaceX API and the SpaceX Wikipedia page. This dual-source approach allowed us to access a comprehensive dataset for analysis.

❖ **SpaceX API**: We leveraged the SpaceX API to access up-to-date and structured information on rocket launches and landings. By making API requests, we retrieved data on launch outcomes, mission details, and other pertinent information. This real-time data stream provided us with the latest insights into SpaceX's missions.

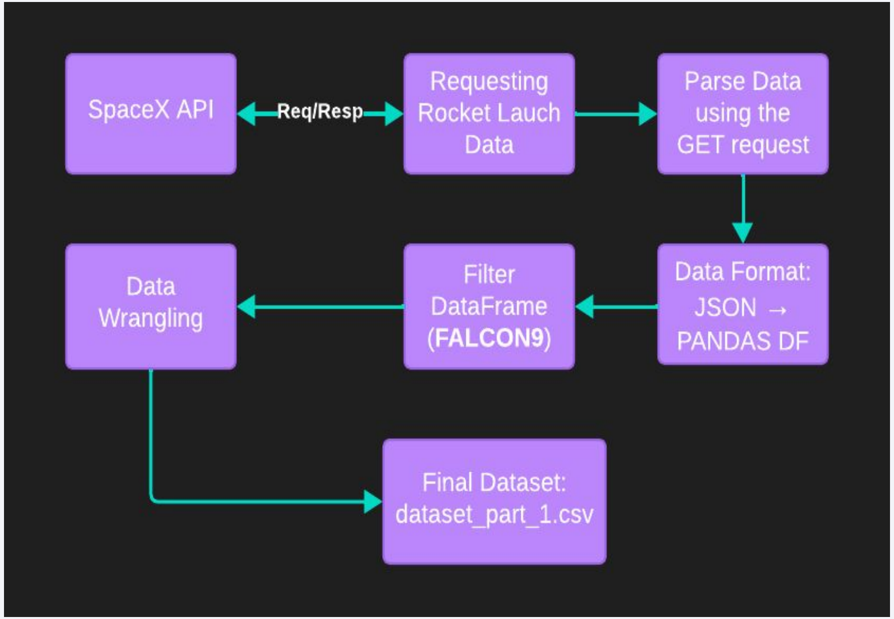
❖ **SpaceX Wikipedia Page**: In parallel, we consulted the SpaceX Wikipedia page, a valuable source for historical records and contextual data. This page offered a wealth of information, including mission descriptions, historical context, and specific details about rocket landings.

Data Collection – SpaceX API

To access data from the SpaceX API, you would typically need to make HTTP requests to the API endpoints, parse the JSON responses, and then use the data for your analysis or applications.

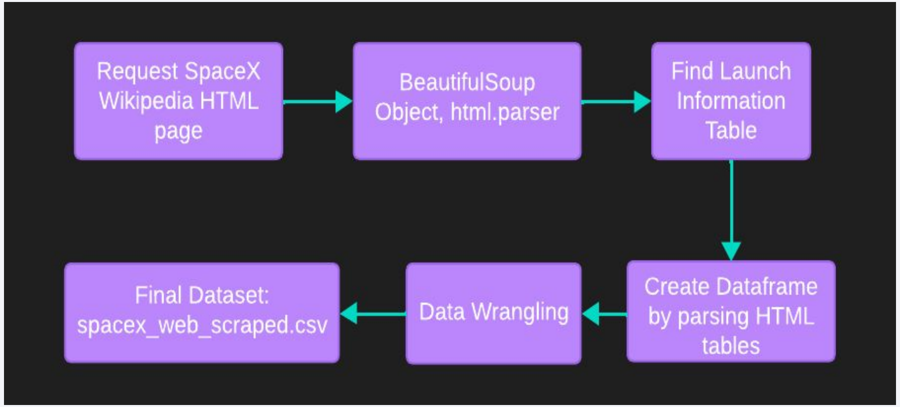
Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude, Class.

GitHub Link



Data Collection - Scraping

Scraping data from a Wikipedia page involves extracting information from the HTML structure of the page using web scraping techniques. I'll provide a general outline of the steps involved in scraping data from a Wikipedia page.



Data Columns: Flight No., Date, Time, Version booster, Launchsite, Payload, Payload mass, Orbit, Customer, Launch Outcome, Booster landing.

GitHub Link

Data Wrangling

In data wrangling process, we encountered diverse mission outcomes related to SpaceX booster landings. These outcomes included cases where landing attempts were made but did not result in a successful landing. To simplify an analysis and prepare the data for predictive modeling, we transformed these outcomes into training labels with a binary classification:

1 was assigned to indicate that the booster successfully landed.

0 was assigned to indicate that the landing was unsuccessful.

- ❑ True Ocean: Assigned 1 for successful ocean landings, 0 for unsuccessful.
- ❑ True RTLS: Assigned 1 for successful ground pad landings, 0 for unsuccessful.
- ❑ True ASDS: Assigned 1 for successful drone ship landings, 0 for unsuccessful.

This simplification streamlines our analysis for accurate predictions.

GitHub Link

EDA with Data Visualization

EDA was conducted on various variables to assess their relationships and suitability for machine learning model training.

Plots Used:

1. Flight Number vs. Payload Mass: Scatter plot to visualize any patterns or trends.
2. Flight Number vs. Launch Site: Bar plot to compare launch sites and their frequency.
3. Payload Mass vs. Launch Site: Scatter plot to explore payload masses by launch site.
4. Orbit vs. Success Rate: Bar plot depicting the success rate of different orbits.
5. Flight Number vs. Orbit: Line chart illustrating the distribution of missions across orbits.
6. Payload vs. Orbit: Scatter plot to analyze payload masses across different orbits.
7. Success Yearly Trend: Line chart showcasing the yearly trend of mission success rates.

Insights:

1. Flight Number vs. Payload Mass and Flight Number vs. Orbit revealed potential trends over time.
2. Launch Site and Payload Mass vs. Launch Site helped identify launch site preferences.
3. Orbit vs. Success Rate provided insights into mission success rates.
4. Yearly trend analysis indicated fluctuations in success rates over time.

GitHub Link

EDA with SQL

Process:

1. **Data Loading**: We successfully loaded the dataset into an IBM DB2 Database for efficient data management and analysis.
2. **SQL Queries**: Leveraging SQL and Python integration, we executed queries to gain deeper insights into the dataset.

Key Queries:

1. **Launch Site Names**: We queried launch site names to understand the geographical distribution of SpaceX launches.
2. **Mission Outcomes**: Queries were made to gather information about mission outcomes, helping us assess the success rate of SpaceX missions.
3. **Payload Sizes**: We queried various payload sizes of customers, providing insights into payload distribution and demand.
4. **Booster Versions**: Queries were executed to identify different booster versions used in SpaceX missions, aiding in booster performance analysis.
5. **Landing Outcomes**: Information about landing outcomes was extracted through queries, enabling us to analyze the success of booster landings.

GitHub Link

Build an Interactive Map with Folium

Description: We utilized Folium, a Python library, to create an interactive map that enhances our understanding of SpaceX launch sites and their proximity to key locations.

Map Features:

- Launch Sites: Marked on the map to show their geographical distribution.
- Successful and Unsuccessful Landings: Visualized to understand the outcomes of SpaceX missions in various locations.
- Proximity to Key Locations: Demonstrated with examples of proximity to Railway, Highway, Coast, and City.

Purpose:

- This interactive map allows us to explore why SpaceX launch sites are strategically located.
- It offers a visual representation of successful landings relative to specific geographical locations.

Impact:

- This map aids decision-making and site selection for future SpaceX missions.
- It provides a clear visual of the success and impact of SpaceX's operations in different areas.

GitHub Link

Build a Dashboard with Plotly Dash

Description: We developed an interactive dashboard using Plotly Dash, offering dynamic insights into SpaceX launch success rates, payload masses, and booster versions.

Dashboard Components:

- Pie Chart: Allows users to select from two options: Distribution of successful landings across all launch sites. Individual launch site success rates.
- Scatter Plot: Provides two input options:(a) All launch sites or individual sites
  (b) Payload mass selection within the range of 0 to 10,000 kg.

Purpose:

- The Pie Chart visualizes overall launch site success rates.
- The Scatter Plot helps analyze success variations across launch sites, payload masses, and booster versions.

Benefits:

Enhanced interactivity for data exploration and decision support. Efficient visualization of launch success patterns and trends.

Impact: This dashboard assists in making informed decisions regarding launch site selection, payload handling, and booster version categorization.

It provides actionable insights into SpaceX mission planning and operations.

GitHub Link

Predictive Analysis (Classification)

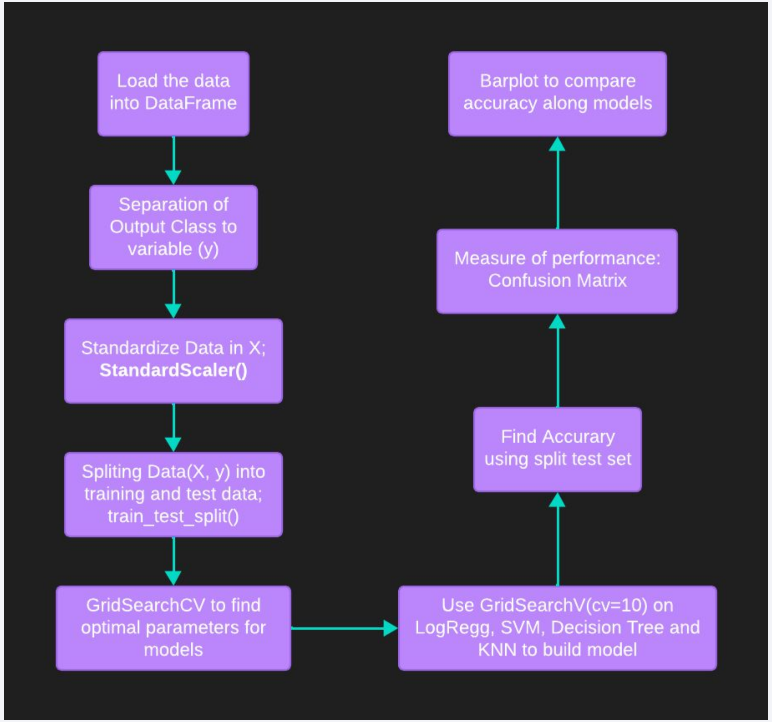
Description: We conducted predictive analysis using classification techniques to forecast the success or failure of SpaceX rocket landings.

Purpose: Predictive analysis aids in anticipating mission outcomes, enhancing mission planning, and minimizing risks.

Benefits:

- Informed decision-making in real-time, allowing SpaceX to optimize launch strategies.
- Improved safety and cost-effectiveness through accurate landing predictions.

Impact: Predictive modeling contributes to the success and sustainability of SpaceX missions, ensuring safe and efficient rocket landings.



GitHub Link

Results

Key Findings: I achieved notable results through extensive data analysis, machine learning, and predictive modeling.
Model Accuracy: All machine learning models (Logistic Regression, Support Vector Machine, Decision Tree Classifier, K Nearest Neighbors) consistently delivered an accuracy rate of approximately 83.33%.
Over-Prediction: All models tended to over-predict successful landings, highlighting the need for additional data for more accurate predictions.
Recommendations: Gather more data to address over-prediction and improve model determination. Continue refining models and exploring advanced machine learning techniques for enhanced accuracy.
Impact: The project lays the foundation for informed decision-making in SpaceX missions, enhancing safety and mission success.
Future Work:

- Further research and data collection to refine models.
- Exploration of advanced machine learning algorithms and techniques.
- Ongoing collaboration to ensure the continued success of space exploration missions.

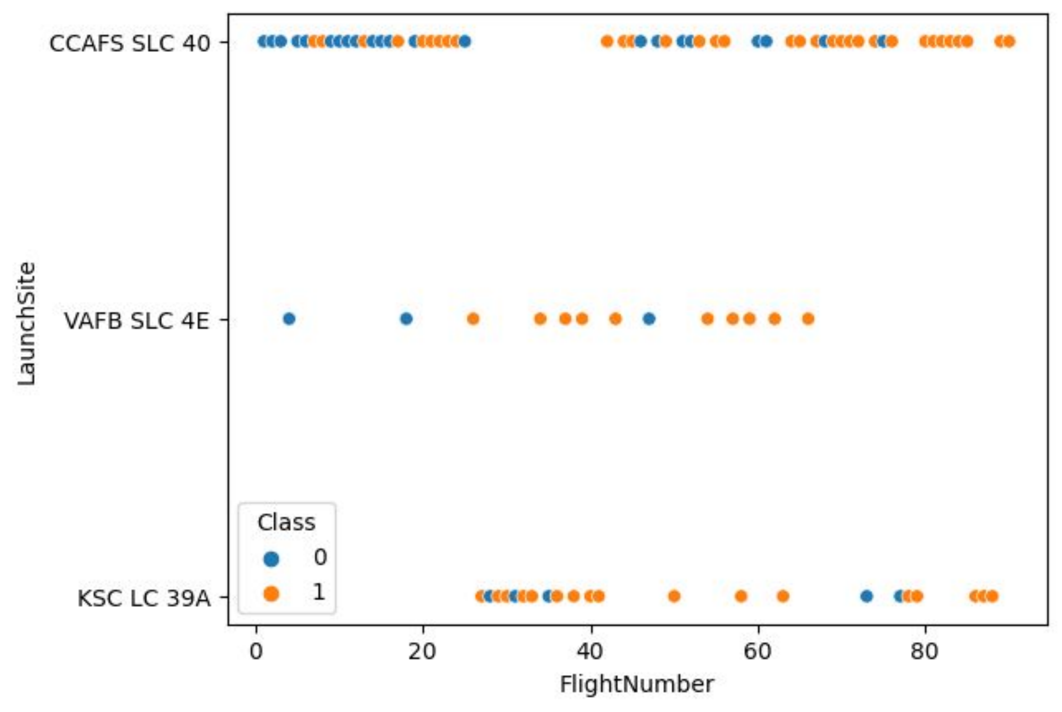
Section
2

Insights drawn from EDA

Flight Number vs. Launch Site

Orange indicates successful launch(Class 1)
Blue indicates unsuccessful launch.(Class 0)

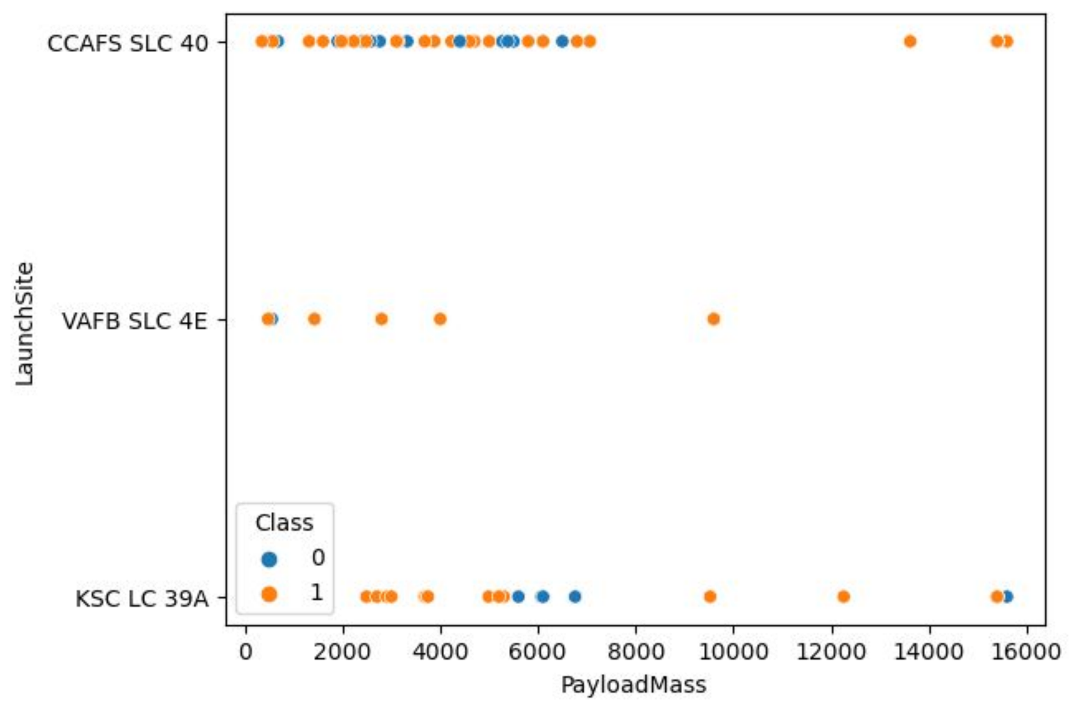
- ❑ Graphic suggests an increase in success rate over time (indicated in Flight Number).
- ❑ Likely a big breakthrough around flight 20 which significantly increased success rate. “***CCAFS SLC 40***” appears to be the main launch site as it has the most volume.



Payload vs. Launch Site

Orange indicates successful launch(Class 1)
Blue indicates unsuccessful launch.(Class 0)

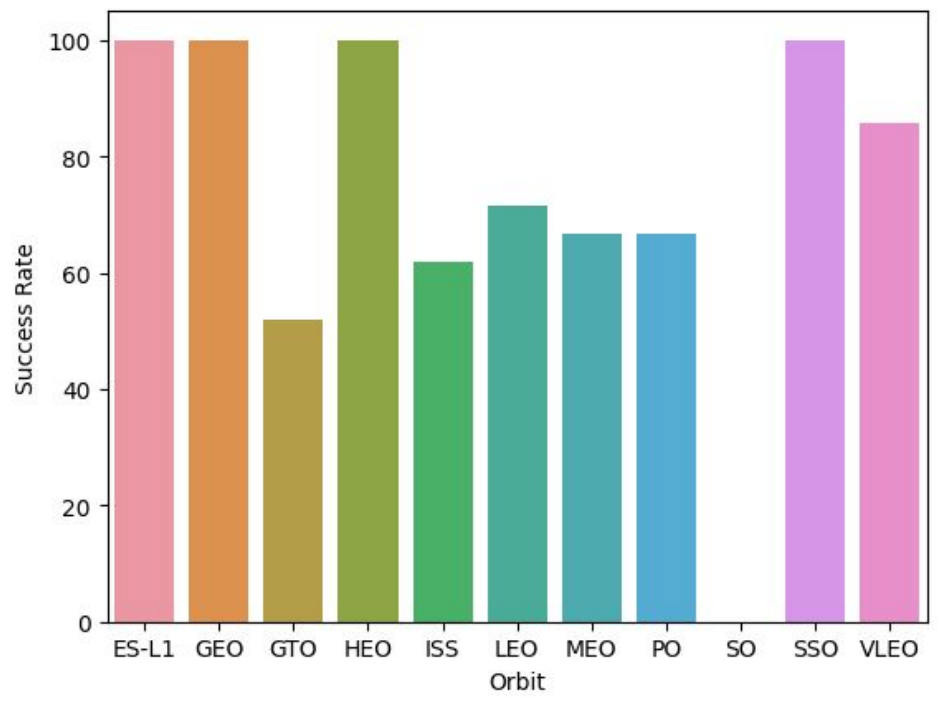
- ❏ Payload mass appears to fall mostly between 0-7000 kg. Different launch sites also seem to use different payload mass.
- ❏ The **VAFB-SLC** launchsite there are no rockets launched for heavy payloadMass(greater than 10000).



Success Rate vs. Orbit Type

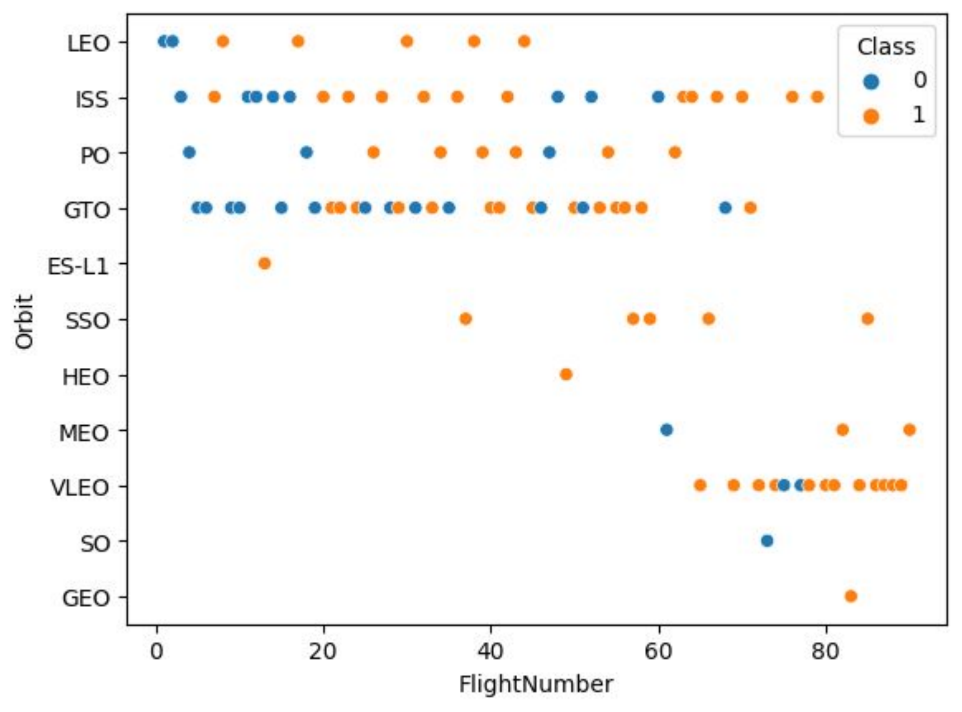
1. ES-L1 (1): 100% success rate
2. GEO (1): 100% success rate
3. HEO (1): 100% success rate
4. SSO (5): 100% success rate
5. VLEO (14): Decent success rate with attempts
6. SO (1): 0% success rate
7. GTO (27): Around 50% success rate, largest sample

- **ES-L1**, **GEO**, **HEO**, and **SSO** launch site categories have a perfect 100% success rate, albeit with small sample sizes.
- **VLEO** demonstrates a decent success rate while actively attempting missions with a larger sample size.
- **GTO**, with the largest sample size of 27, maintains an approximately 50% success rate.



Flight Number vs. Orbit Type

- Orange indicates successful launch(Class 1)
- Blue indicates unsuccessful launch.(Class 0)



- ❑ Over time, SpaceX's launch orbit preferences have evolved significantly, correlating with mission outcomes.
- ❑ Initially, SpaceX predominantly focused on Low Earth Orbit (**LEO**) missions, achieving moderate success rates.
- ❑ In recent launches, there has been a shift back to Very Low Earth Orbit (**VLEO**).

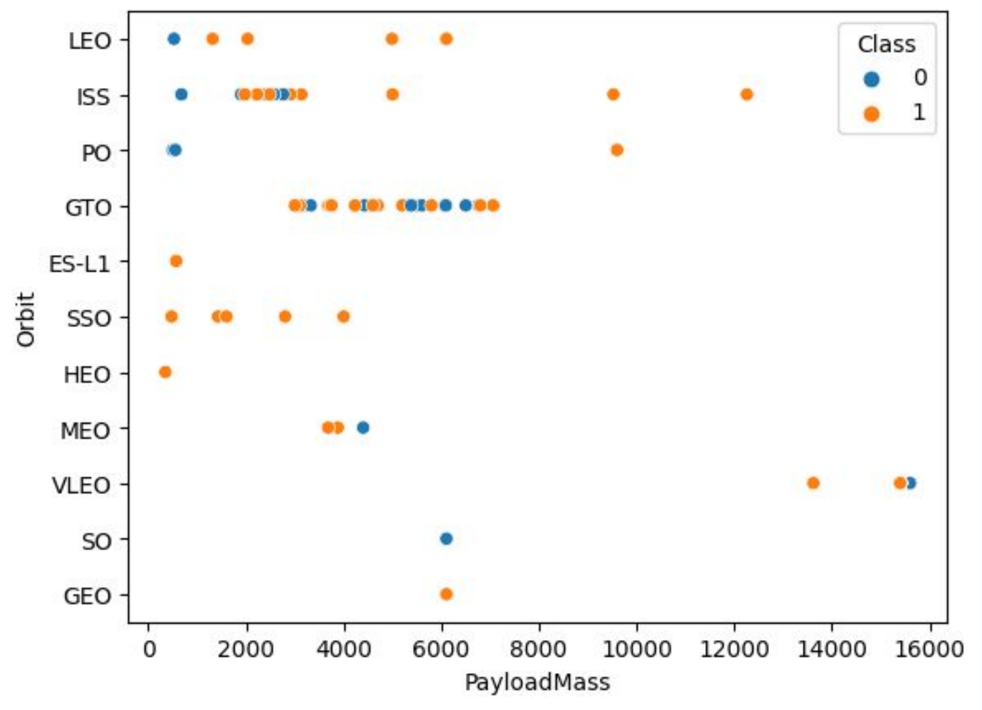
Payload vs. Orbit Type

- Orange indicates successful launch(Class 1)
- Blue indicates unsuccessful launch.(Class 0)

❑ Low Earth Orbit (**LEO**) and Sun-Synchronous Orbit (**SSO**): These orbits tend to have relatively low payload masses.

❑ Very Low Earth Orbit (**VLEO**): This orbit type primarily accommodates higher-end payload mass values.

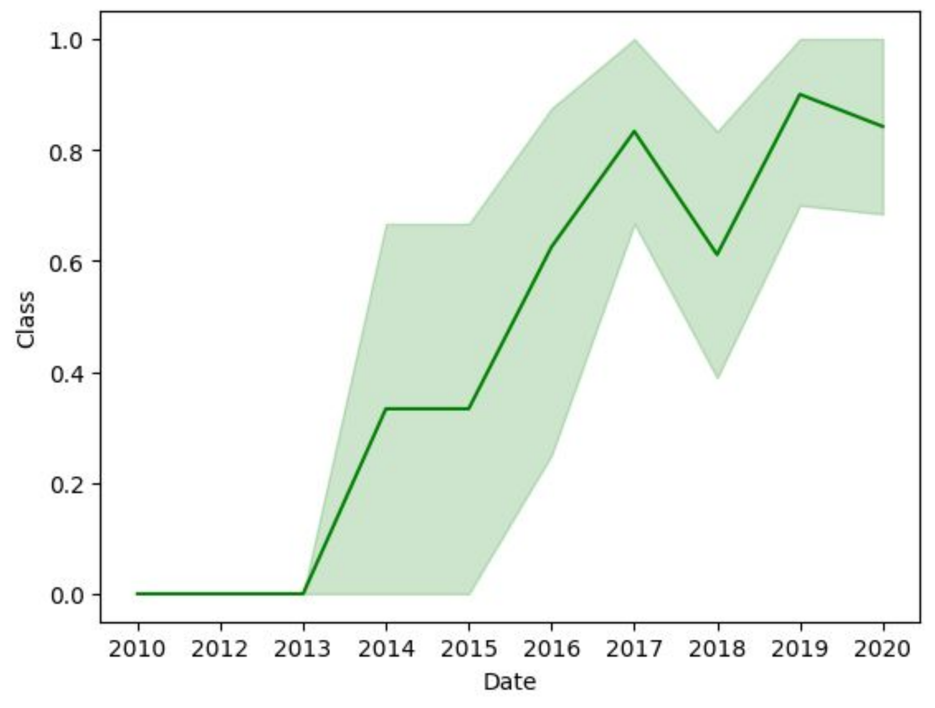
❑ Geostationary Transfer Orbit (**GTO**): Most GTO launches fall within the payload mass range of 3000 to 8000.



Launch Success Yearly Trend

- Since 2013, success rates have generally shown an increasing trend
- A slight dip in success rates was observed in 2018.
- In recent years, success rates have stabilized at around 80%.

This trend informs future mission planning and risk assessment, contributing to mission success and overall success in space exploration endeavors.



All Launch Site Names

```
In [8]: %sql select DISTINCT LAUNCH_SITE from SPACEXTABLE

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

Out[8]:

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

Unique launch site names used by SpaceX:

1. “**CCAFS LC 40**” refers to Cape Canaveral Air Force Station Launch Complex 40
2. “**CCAFS SLC 40**” stands for Cape Canaveral Air Force Station Space Launch Complex 40
3. “**KSC LC-39A**” stands for Kennedy Space Center Launch Complex 39A
4. “**VAFB SLC-4E**” refers to Vandenberg Space Force Base Space Launch Complex 4E

Launch Site Names Begin with 'CCA'

```
In [9]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5

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

Out[9]:

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

First five entries in database with Launch Site name beginning with “CCA”

Total Payload Mass

In [10]:
```
%sql select sum(payload_mass__kg_) as sum from SPACEXTABLE
where customer like 'NASA (CRS)'
```

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

Out[10]:

sum
45596

- The sum of total payload mass in kilograms where NASA was the customer: **45596** kg
- This query filters the dataset to select records where NASA was the customer and the mission was part of the CRS(Commercial Resupply Services) program. It will then calculate the sum of the payload mass in kilograms for these selected records.

Average Payload Mass by F9 v1.1

```
In [11]: %sql select avg(payload_mass__kg_) as Average from SPACEXTABLE
         where booster_version like 'F9 v1.1%'

 * sqlite:///my_data1.db
Done.

Out[11]:
```

Average
2534.6666666666665

The query calculates the average payload mass for SpaceX missions with booster versions starting with 'F9 v1.1' which is **2534.66** kg, providing insights into the typical payload capacity of these missions.

First Successful Ground Landing Date

```
In [12]: %sql select min(date) as Date from SPACEXTABLE
         where mission_outcome like 'Success'

          * sqlite:///my_data1.db
         Done.

Out[12]:
```

Date
2010-04-06

the query finds the earliest date among missions with a successful outcome in the "SPACEXTABLE" dataset, providing information about the first successful mission date in the dataset which is **6th April, 2010**

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [13]: %sql select Booster_Version from SPACEXTABLE
         where (Mission_Outcome like 'Success')
         AND
         (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
         AND
         (Landing_Outcome like 'Success (drone ship)')

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

Out[13]:

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

This query retrieves the "Booster_Version" for missions meeting these specific criteria, providing information about successful missions within the specified payload mass range that also had successful drone ship landings.

- The "Mission_Outcome" is 'Success'
- The PAYLOAD_MASS__KG_" is between 4000 and 6000 kilograms.
- The "Landing_Outcome" is 'Success (drone ship).'

Total Number of Successful and Failure Mission Outcomes

In [14]:
```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTABLE
GROUP BY mission_outcome
ORDER BY mission_outcome
```

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

Out[14]:

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

This query retrieves list of total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
In [15]: %sql select booster_version from SPACEXTABLE
         where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTABLE)

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

Out[15]:

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

The query identifies the "Booster_Version" used in missions with the maximum payload mass recorded in the "SPACEXTABLE" dataset.

2015 Launch Records

In [58]:
```
%sql select substr(DATE,6,2) as Month, landing_outcome, booster_version, launch_site FROM SPACEXTABLE
where (DATE like '2015%')
AND
(landing_outcome like "Failure%)")
```

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

Out[58]:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

The query retrieves information about missions that occurred in the year 2015, had a landing outcome starting with "Failure," and displays the month, landing outcome, booster version, and launch site for these missions.

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

In [72]:
```
%sql SELECT Landing_Outcome, COUNT(*) AS COUNT FROM SPACEXTABLE
WHERE DATE BETWEEN "2010-06-04" AND "2017-03-20"
GROUP by Landing_Outcome
ORDER BY COUNT DESC
```

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

Out[72]:

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

The query retrieves the count of each unique "Landing_Outcome" within a specified date range and orders the results by count in descending order.

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

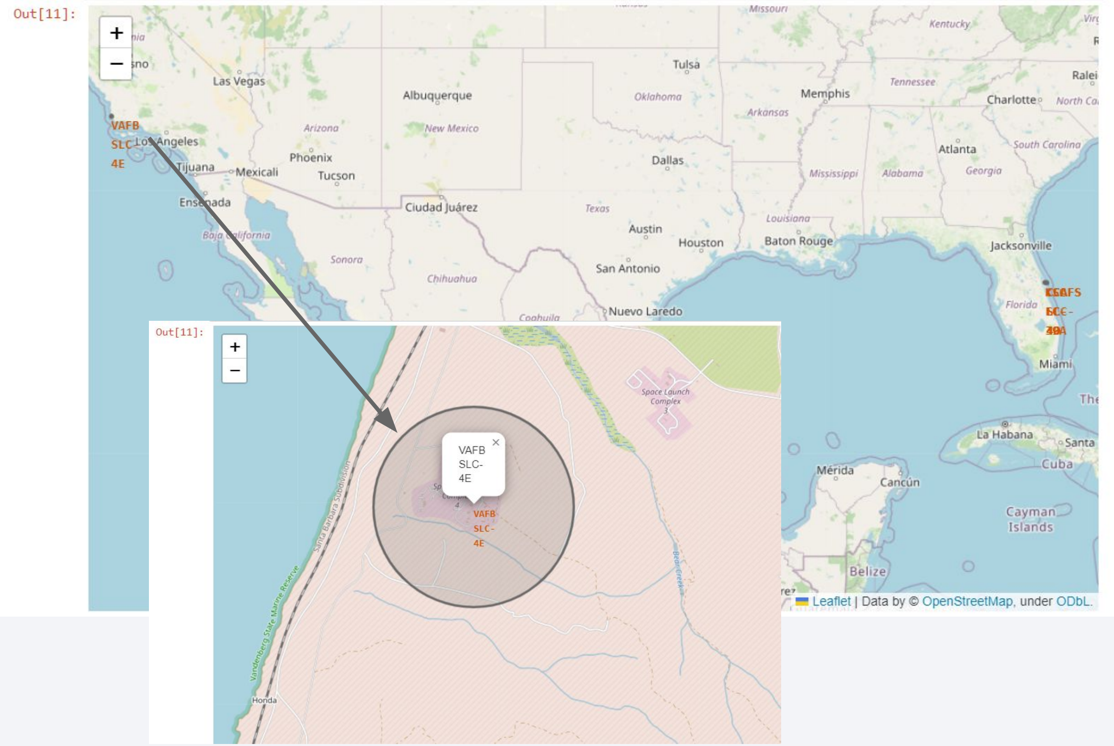
Big Map:

- Displaying all SpaceX launch sites relative to a U.S. map.
- Highlights the proximity of launch sites to the ocean.

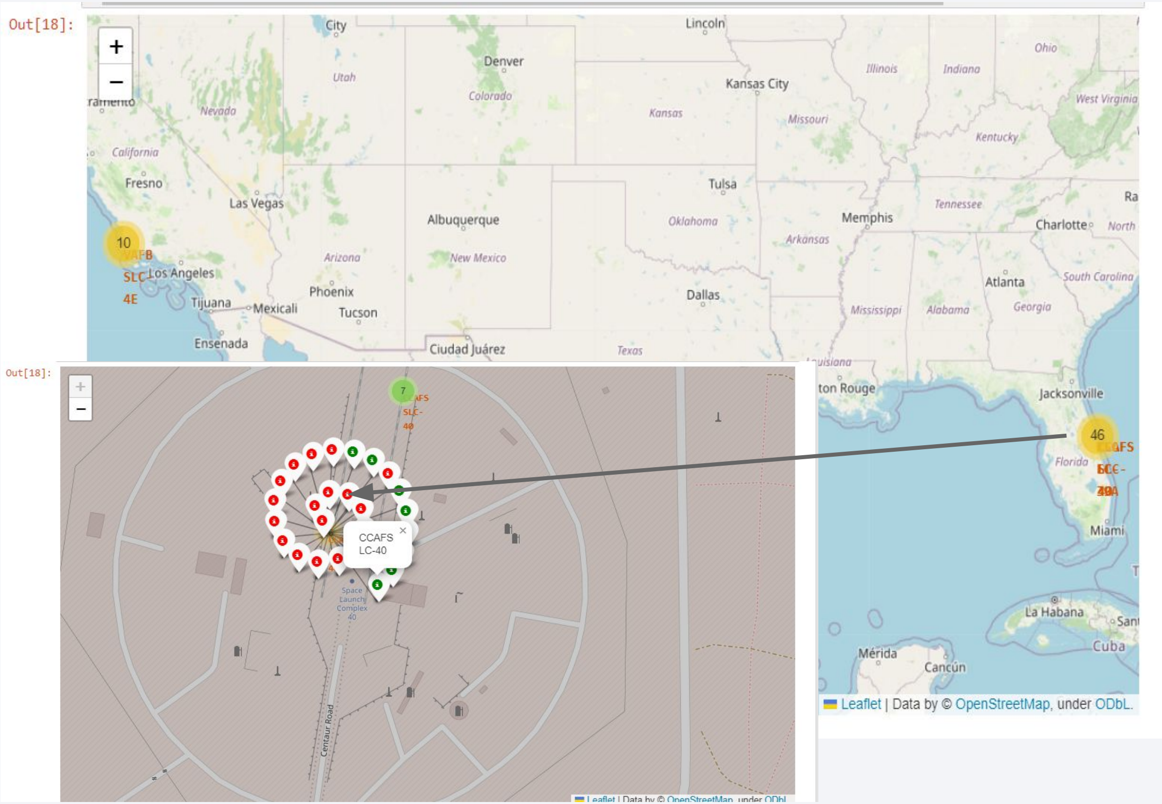
Magnified Map:

- Focusing on the specific launch site "VAFB SLC-4E" located near Los Angeles.
- Demonstrates the coastal location of launch facilities.

Coastal launch sites provide easy access to a wide range of orbital inclinations.



Color-Labeled Launch Outcomes

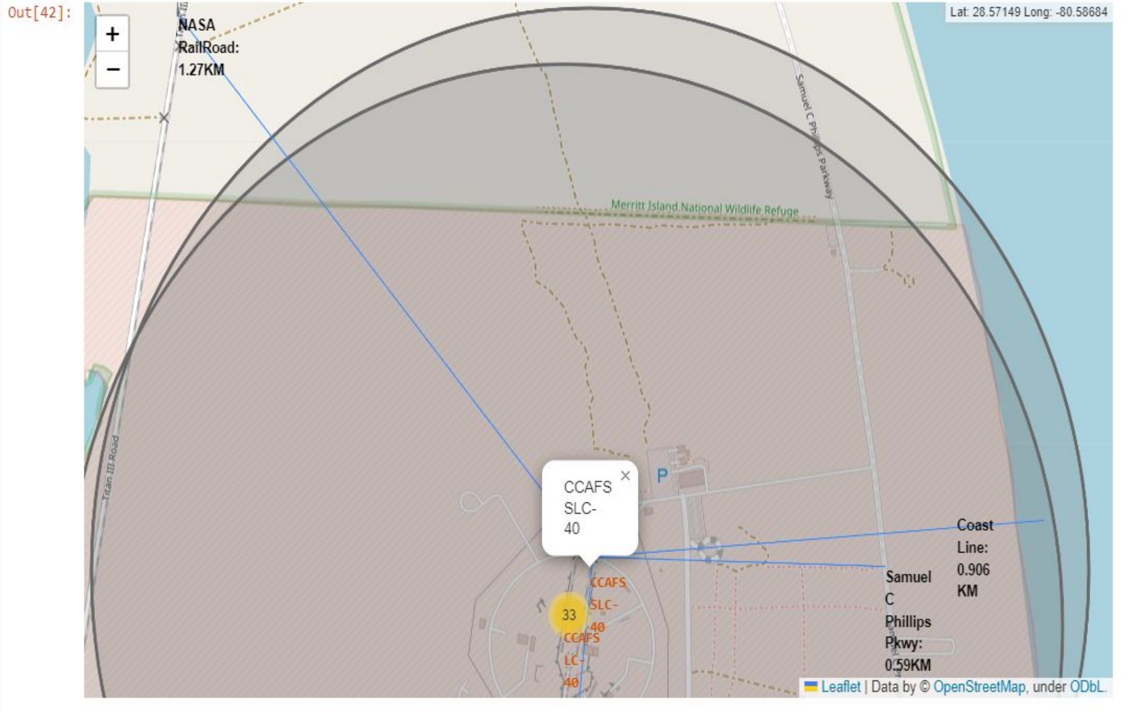


Green Icon: Successful landings.
Red Icon: Failed landings.

Example: CCAFS LC-40
The magnified map of CCAFS LC-40 shows **7** successful landings (green icon) and **19** failed landings (red icon).

These color-coded maps offer a user-friendly way to assess the success and performance of SpaceX launch sites, assisting in data-driven decision-making for future missions.

Key Location Proximities



Utilizing "**CCAFS SLC-40**" as an example, we explore the proximity of launch sites to critical transportation and safety features.

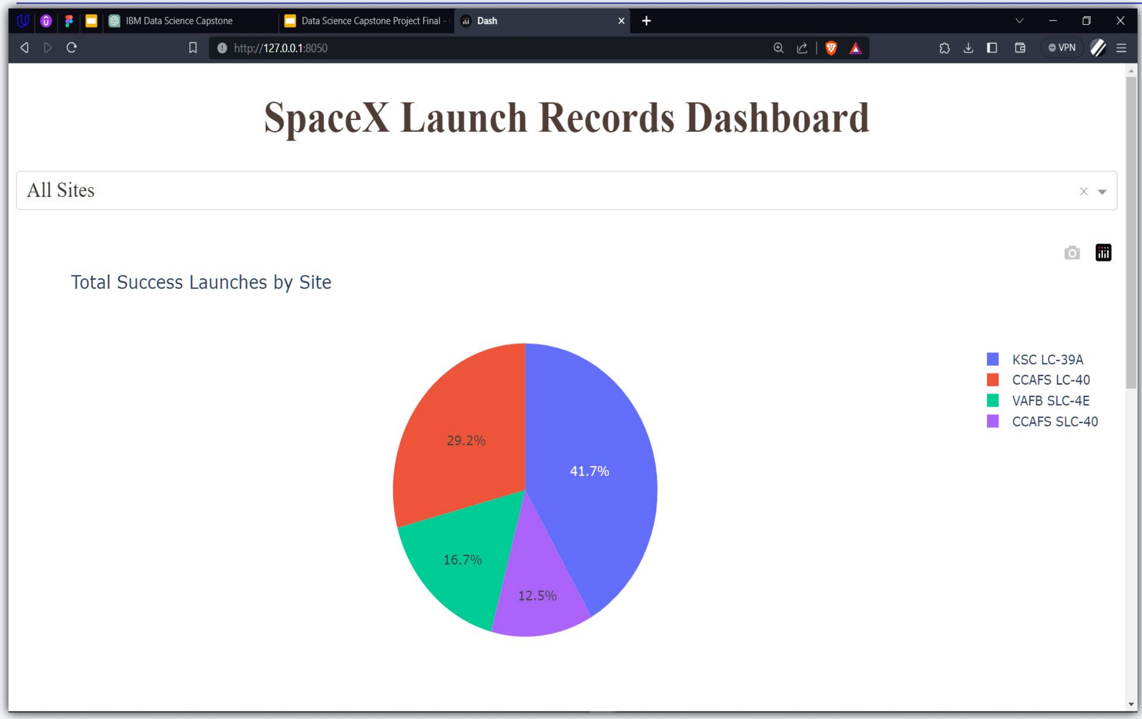
- ❑ Coast Line (East): **0.91 km**
- ❑ City (Melbourne): **51.76 km**
- ❑ Railway (NASA Rail Road): **1.27 km**
- ❑ Highway (Samuel C Phillips Pkwy): **0.59 km**

- Proximity to railways, highways, and cities aids in efficient payload transport and launch operations.

Section
4

Build a Dashboard with Plotly Dash

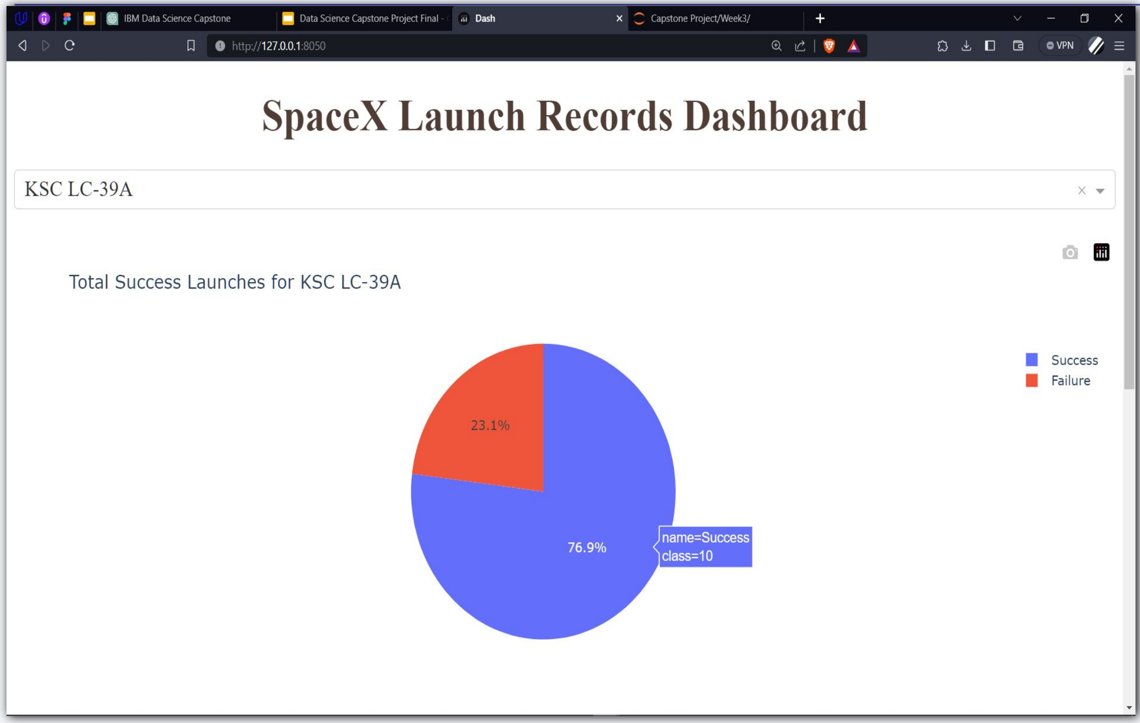
Successful Launches by Launch Sites



Pie Chart: Represents the distribution of successful landings across all SpaceX launch sites.

- ❑ KSC LC-39A has the highest percentage of successful landings, comprising **41.7%**.
- ❑ CCAFS LC 40 and VAFB SLC-4E also show significant success rates at **29.2%** and **16.7%**, respectively.
- ❑ CCAFS SLC 40 contributes with a **12.5%** success rate.

Highest launch success ratio Launch Site



KSC LC-39A: This launch site stands out with the highest success rate, marked in Blue, featuring **10** *successful* landings and **3** *failed* landings.

- KSC LC-39A demonstrates a strong track record, making it a strategic choice for future missions.
- This distribution provides a clear overview, highlighting KSC LC-39A as a top-performing site.

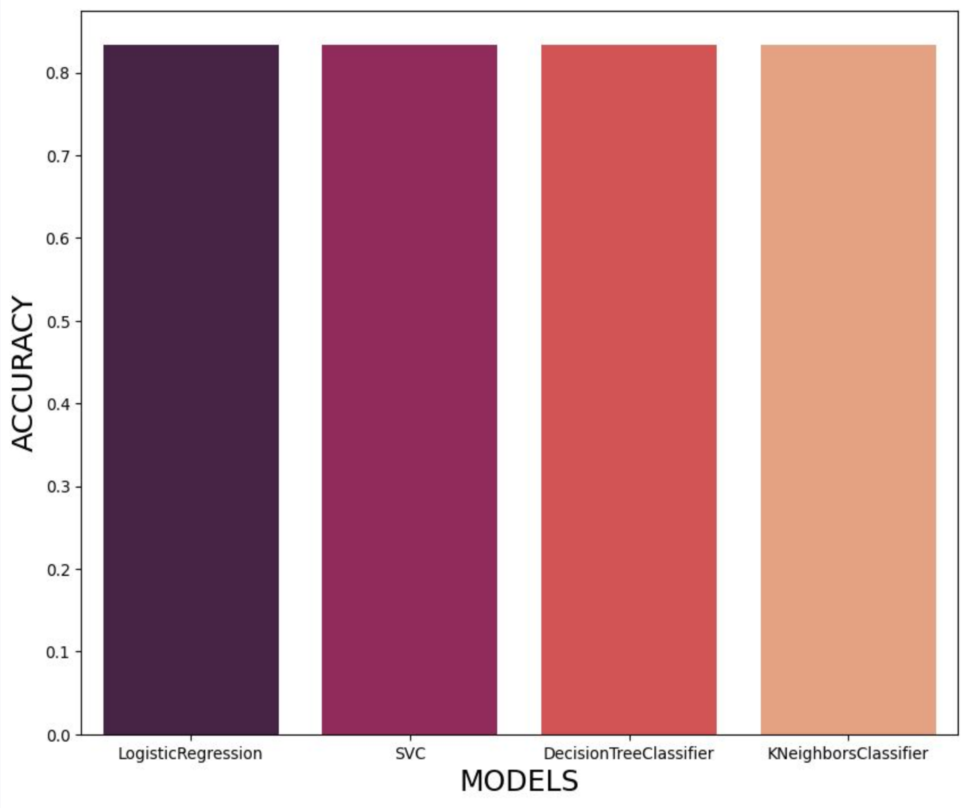
Payload Mass vs. Success vs. Booster Version Category

- ❑ Payload Range Selector: Allows users to select the payload mass range(0 kg to 10,000 kg).
- ❑ Class Indicator: Classifies successful landings as "1" and failures as "0."
- ❑ Scatter Plot: Displays the payload mass on the x-axis, success/failure on the y-axis, booster version category in color, and the number of launches represented by point size.

Section
5

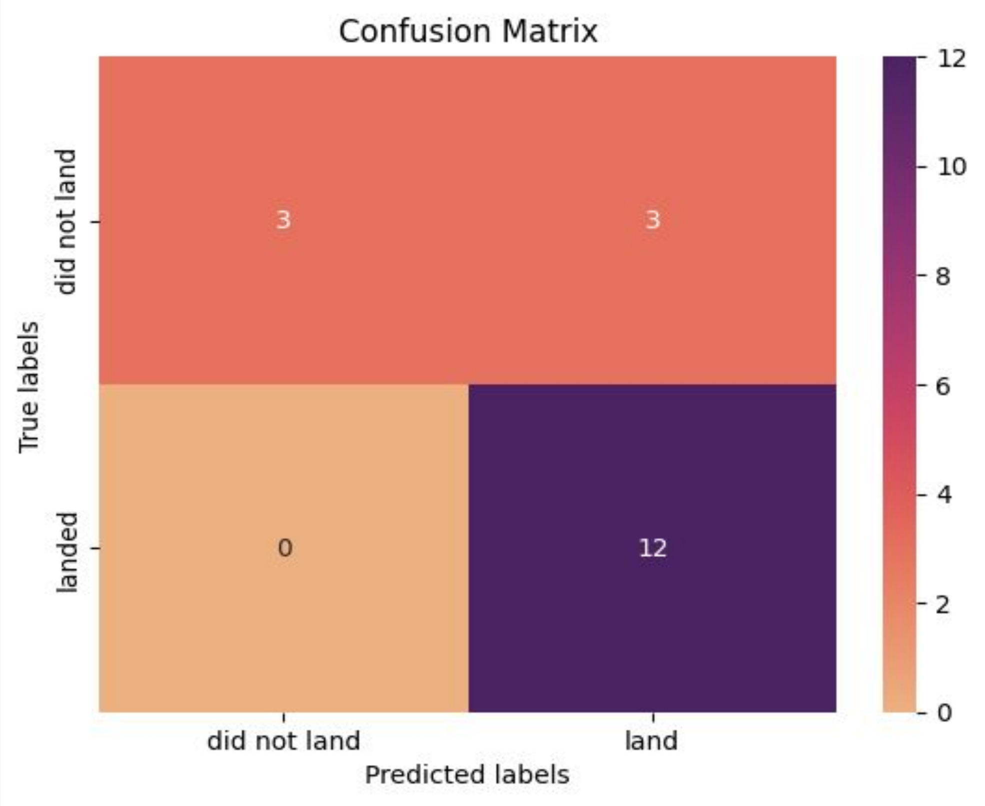
Predictive Analysis (Classification)

Classification Accuracy



- Model Comparison: All machine learning models produced virtually identical accuracy on the test set, achieving **83.33%** accuracy.
- While the current results are promising, expanding the dataset can help provide more robust and reliable insights into model performance.
- This slide emphasizes the accuracy achieved in the current dataset while acknowledging the potential impact of small test sizes on variability and the need for additional data for more confident model assessments.

Confusion Matrix



- **True Positives** (TP): There are 3 cases where the model correctly predicted positive outcomes.
- **True Negatives** (TN): There are 12 cases where the model correctly predicted negative outcomes.
- **False Positives** (FP): There are 3 cases where the model incorrectly predicted positive outcomes when the actual outcome was negative.
- **False Negatives** (FN): There are no cases where the model incorrectly predicted negative outcomes when the actual outcome was positive.

Conclusions

1. **Data Collection and Wrangling**: The project began with data collection from public sources, including SpaceX APIs and Wikipedia. Data wrangling involved cleaning, transforming, and preparing the dataset for analysis.
2. **Exploratory Data Analysis (EDA)**: EDA revealed trends in launch success over time, payload mass correlations with orbit types, and the significance of launch site locations.
3. **Visualization and Interactive Tools**: The project incorporated Folium maps, Plotly Dashboards, and charts to visually represent the data, enhancing our understanding of launch site locations and success factors.
4. **Predictive Analysis**: Four machine learning models were developed, showing consistent accuracy of 83.33% on a small test set. However, the small dataset size suggests the need for more data to determine the best model.
5. **Strategic Insights**: The analysis highlighted the importance of coastal launch sites, proximity to transportation infrastructure, and the reliability of KSC LC-39A.
6. **Future Directions**: To improve model determination and accuracy, future work should focus on gathering additional data and exploring advanced machine learning techniques.

This project provides valuable insights into SpaceX's launch operations, laying the foundation for data-driven decisions and strategic planning in space exploration.

Appendix

GITHUB REPOSITORY LINK:

https://github.com/ganeshs14/AppliedDSCapstone/tree/main

Thanks to all Instructors:
Dr. Pooja, Romeo Kienzler, Joseph Santarcangelo, Polong Lin, Alex Aklson, Rav Ahuja, Saishruthi Swaminathan, Saeed Aghabozorgi, Hima Vasudevan, Azim Hirjani, Aije Egwaikhide, Yan Luo, Levitan Svetlana.

Special Thanks to **Coursera** and **IBM**

Thank you!