IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Farhat Rekaya
11-05-2023

Outline

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

Executive Summary

- Summary of methodologies
  - ➢ Data Collection (API/Web Scraping)
  - ➢ Data wrangling
  - ➢ EDA with SQL and Data Visualization
  - ➢ Machine Learning
- Summary of all results
  - ➢ EDA results
  - ➢ Interactive Dashboards
  - ➢ Predictive Analysis results

Introduction

- Project Background and Context:
  - SpaceX has revolutionized space travel affordability by reusing the first stage of their Falcon 9 rockets. With their launches priced at $62 million, much lower than competitors' costs of over $165 million, predicting the success of the first stage landing becomes crucial in determining launch expenses. Leveraging public information and machine learning models, our goal is to predict whether SpaceX will reuse the first stage.

Introduction

- Questions to be Answered:

  1. How do variables such as payload mass, launch site, number of flights, and orbits influence the success of the first stage landing?
  2. Does the rate of successful landings exhibit an upward trend over the years?
  3. Which algorithm yields the best performance for binary classification in this particular case?

Section 1

Methodology

Methodology

Executive Summary

- Data Collection Methodology:
  - Utilized SpaceX Rest API for data collection
  - Employed Web Scraping techniques from Wikipedia
  - Data Wrangling Process:
  - Filtered the collected data
  - Addressed any missing values
  - Utilized One Hot Encoding for data preparation in binary classification

Methodology

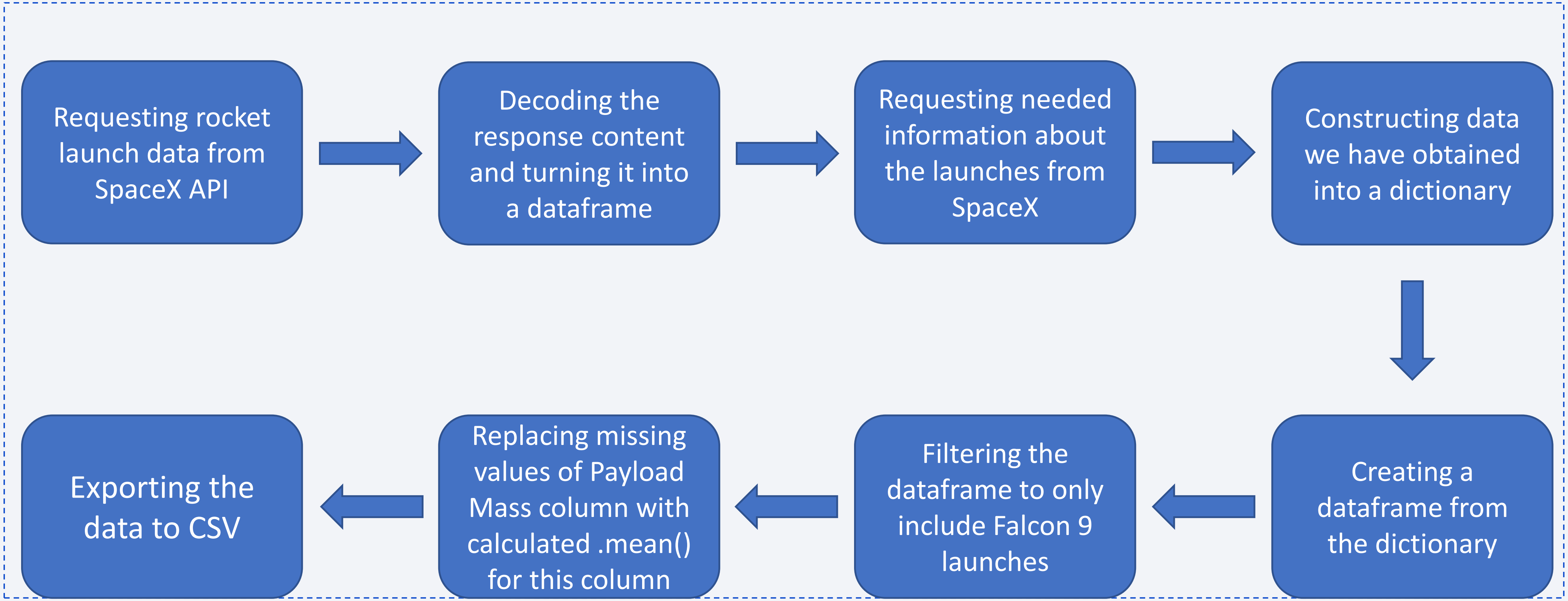
Executive Summary

- Exploratory Data Analysis (EDA):
  - Leveraged visualization and SQL for EDA
  - Interactive Visual Analytics:
  - Employed Folium and Plotly Dash for interactive visual analytics
- Predictive Analysis:
  - Developed, fine-tuned, and evaluated classification models to achieve optimal results.

Data Collection

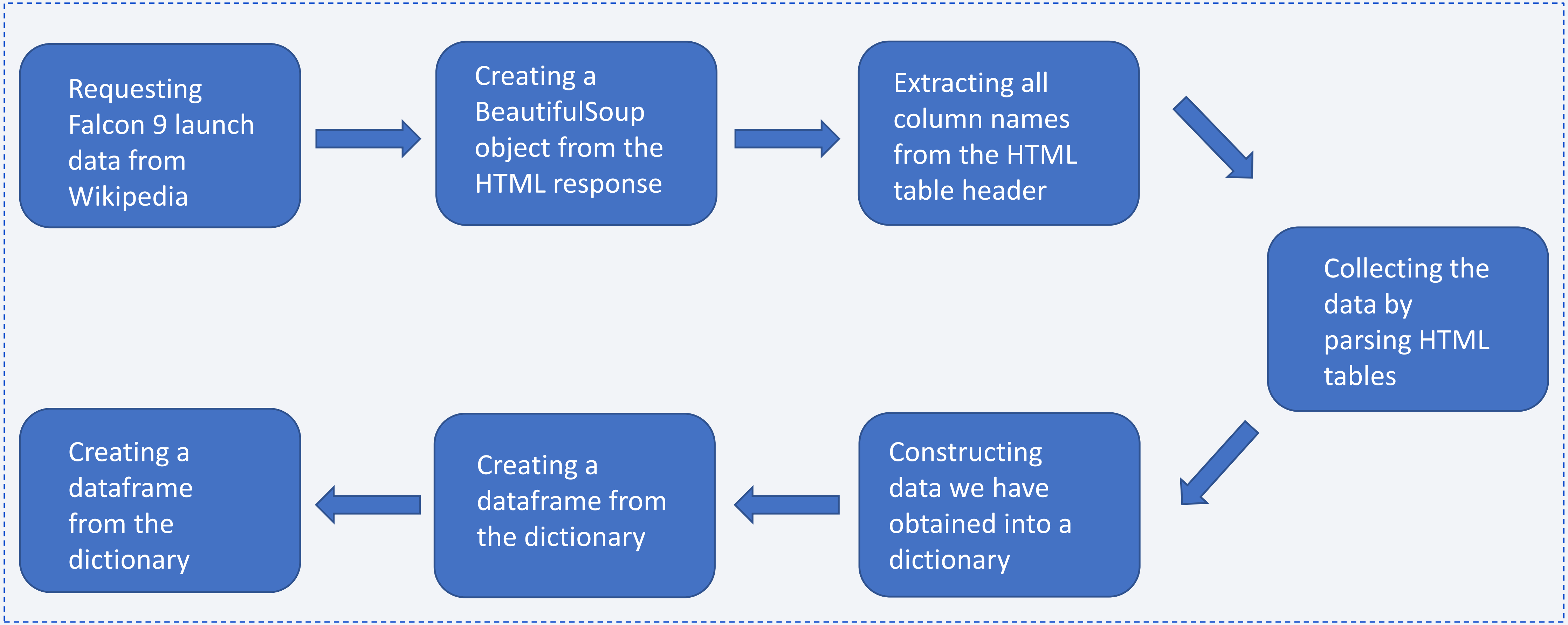
- The data collection process involved utilizing a combination of API requests from the SpaceX REST API and web scraping data from a table found on SpaceX's Wikipedia entry. We employed both of these methods to ensure comprehensive information gathering for a more thorough analysis.
- The following data columns were obtained through the SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- The following data columns were acquired through web scraping from Wikipedia: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API



https://github.com/FARHATREKAYA/IBM_data_science/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

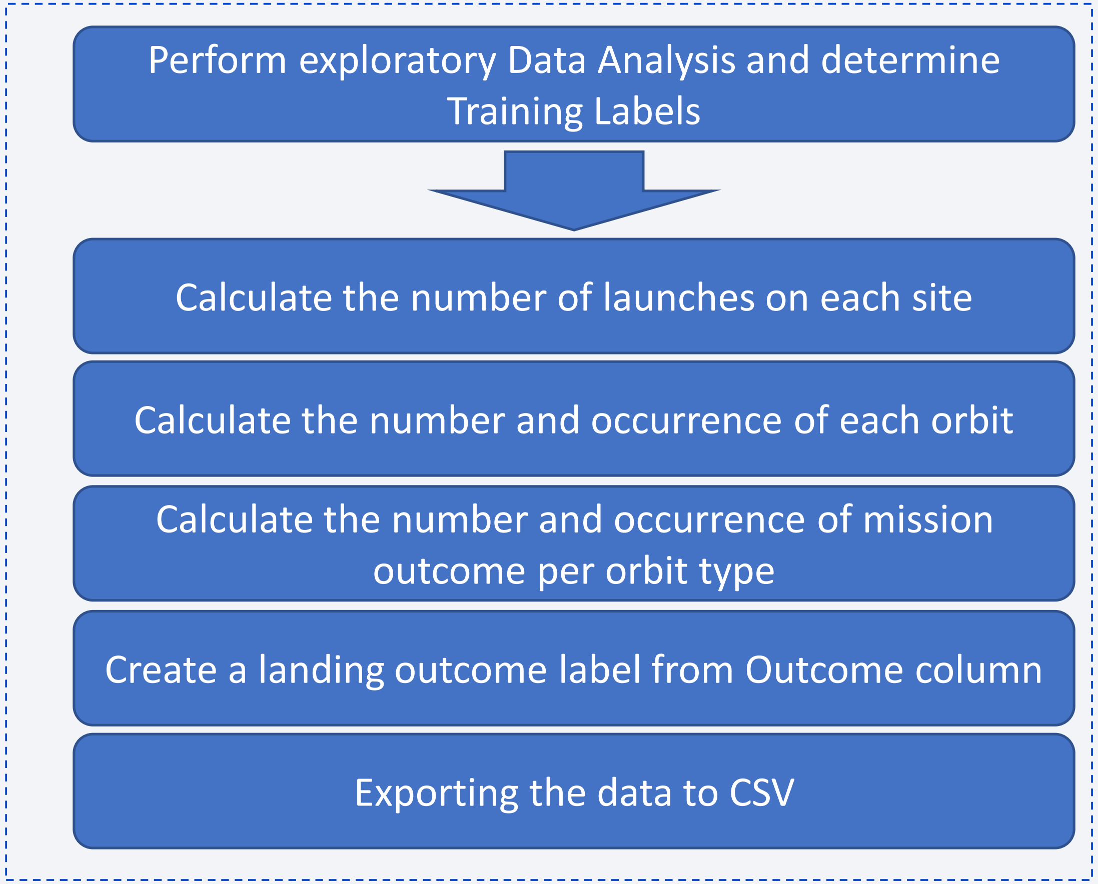
Data Collection - Scraping



https://github.com/FARHATREKAYA/IBM_data_science/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Within the dataset, there are various scenarios where the booster's landing was not successful. In some cases, a landing was attempted but resulted in failure due to accidents. For instance, "True Ocean" indicates a successful mission outcome where the booster landed in a specific region of the ocean, while "False Ocean" signifies an unsuccessful landing in the ocean. Similarly, "True RTLS" denotes a successful landing on a ground pad, whereas "False RTLS" indicates an unsuccessful landing on a ground pad. "True ASDS" signifies a successful landing on a drone ship, whereas "False ASDS" represents an unsuccessful landing on a drone ship.
- To facilitate analysis, we primarily convert these outcomes into training labels. A label of "1" is assigned when the booster successfully lands, while a label of "0" indicates an unsuccessful landing.



https://github.com/FARHATREKAYA/IBM_data_science/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- We generated a series of charts to visualize the data:
  1. Flight Number vs. Payload Mass
  2. Flight Number vs. Launch Site
  3. Payload Mass vs. Launch Site
  4. Orbit Type vs. Success Rate
  5. Flight Number vs. Orbit Type
  6. Payload Mass vs. Orbit Type
  7. Yearly Trend of Success Rate
- Scatter plots were used to examine the relationships between variables. If a relationship was found, these plots could be utilized in machine learning models.
- Bar charts were employed to compare different discrete categories and showcase the relationship between the specific categories being compared and a measured value.
- Line charts were used to demonstrate trends in data over time, particularly in the context of time series analysis.

https://github.com/FARHATREKAYA/IBM_data_science/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Performed SQL queries to:
  1. Retrieve unique launch site names.
  2. Fetch five records with launch sites starting with 'CCA'.
  3. Calculate the total payload mass carried by NASA (CRS) boosters.
  4. Determine the average payload mass for booster version F9 v1.1.
  5. Find the date of the first successful ground pad landing.
  6. List the names of boosters successfully landing on a drone ship with payload mass between 4000 and 6000.
  7. Count the total number of successful and failed mission outcomes.
  8. Identify booster versions with the maximum payload mass.
  9. List failed landing outcomes on a drone ship, including booster versions and launch site names for 2015.
  10. Rank landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

- https://github.com/FARHATREKAYA/IBM_data_science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- For all Launch Sites:
  - Plotted markers with circles, popup labels, and text labels for NASA Johnson Space Center using its latitude and longitude coordinates as the starting location.
  - Plotted markers with circles, popup labels, and text labels for all Launch Sites, displaying their geographical locations and proximity to the Equator and coastlines.
- Launch Outcome Markers:
  - Utilized colored markers, such as green for successful launches and red for failed launches, using Marker Cluster to identify Launch Sites with relatively high success rates.
- Distances to Proximities:
  - Added colored lines to illustrate the distances between a Launch Site (e.g., KSC LC-39A) and its nearby features like railways, highways, coastlines, and the closest city.

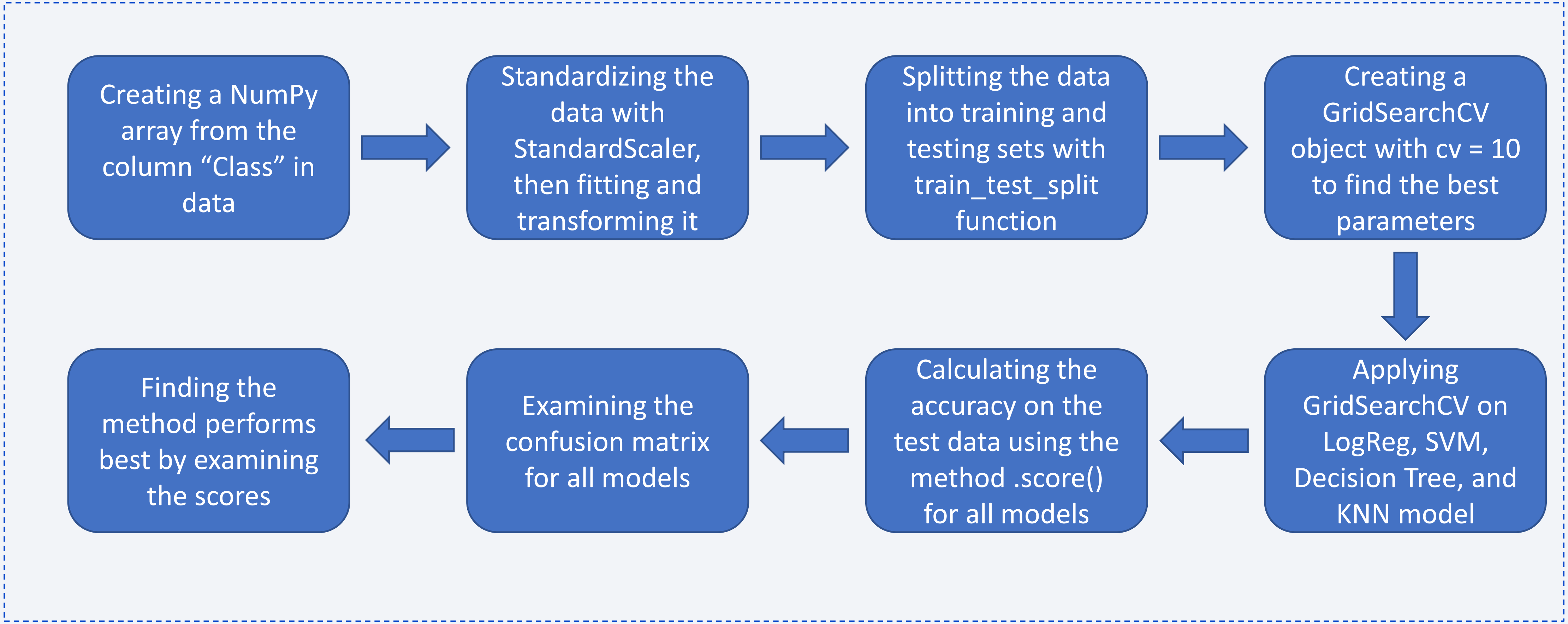
https://github.com/FARHATREKAYA/IBM_data_science/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Implemented a dropdown list feature to facilitate the selection of Launch Sites.
- Pie Chart for Success Launches:
  - Incorporated a pie chart to visualize the total count of successful launches for all Launch Sites. If a specific Launch Site is selected, the chart displays the distribution of Success vs. Failed counts for that particular site.
- Slider for Payload Mass Range:
  - Introduced a slider component that allows users to specify a desired Payload mass range.
- Scatter Chart for Payload Mass vs. Success Rate by Booster Versions:
  - Created a scatter chart to illustrate the relationship between Payload mass and Launch Success rate across different Booster Versions.

https://github.com/FARHATREKAYA/IBM_data_science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/FARHATREKAYA/IBM_data_science/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

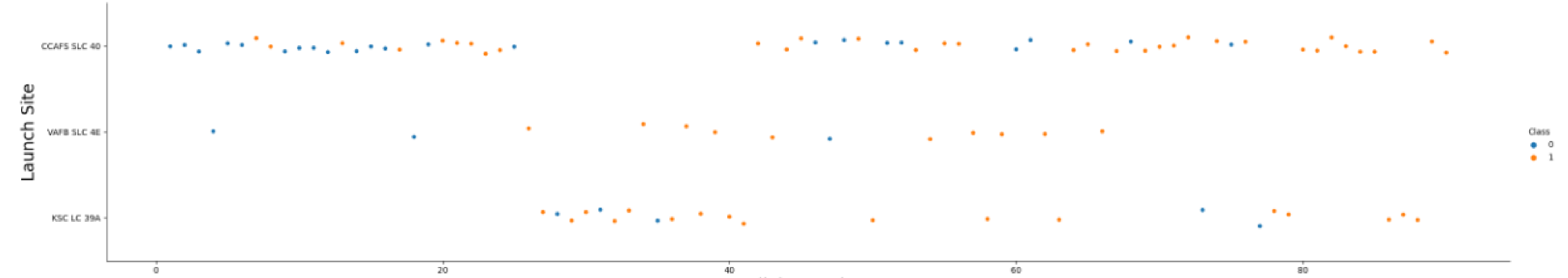
Results

- Exploratory Data Analysis Results:
    - Comprehensive findings and insights derived from the exploratory data analysis process.
- Interactive Analytics Demo in Screenshots:
    - Screenshots showcasing an interactive analytics demo, demonstrating the functionality and features of the analytics platform or tool.
- Predictive Analysis Results:
    - Outcome and findings obtained from predictive analysis, which involve using historical data to make predictions or forecast future outcomes.

Section 2

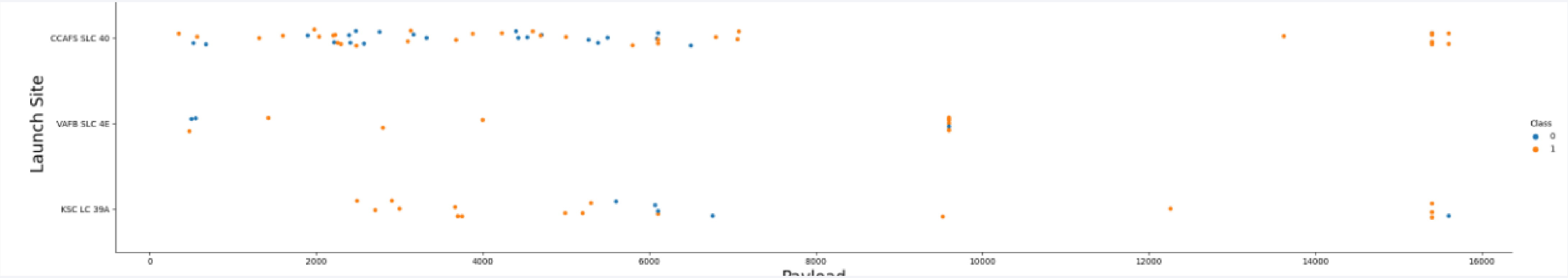
Insights drawn from EDA

Flight Number vs. Launch Site



- Key Observations:
  - The earliest flights in the dataset resulted in failures, while the most recent flights were all successful, suggesting an improvement in mission outcomes over time.
  - The CCAFS SLC 40 launch site accounted for approximately half of all launches, indicating its significance in the space mission.
  - Launch sites VAFB SLC 4E and KSC LC 39A exhibited higher success rates compared to other launch sites, implying favorable conditions or operational efficiency.
  - There is an indication that the success rate of launches tends to improve with each subsequent launch, suggesting a learning curve or continuous optimization efforts.

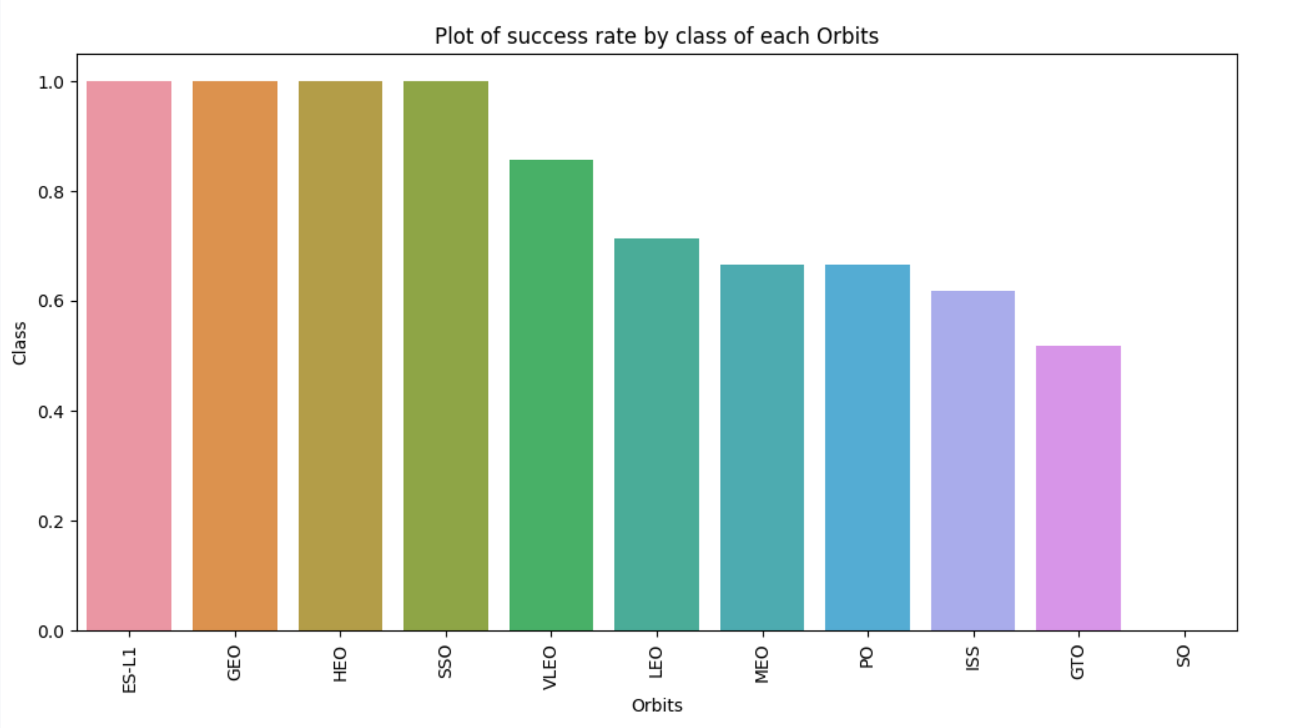
Payload vs. Launch Site



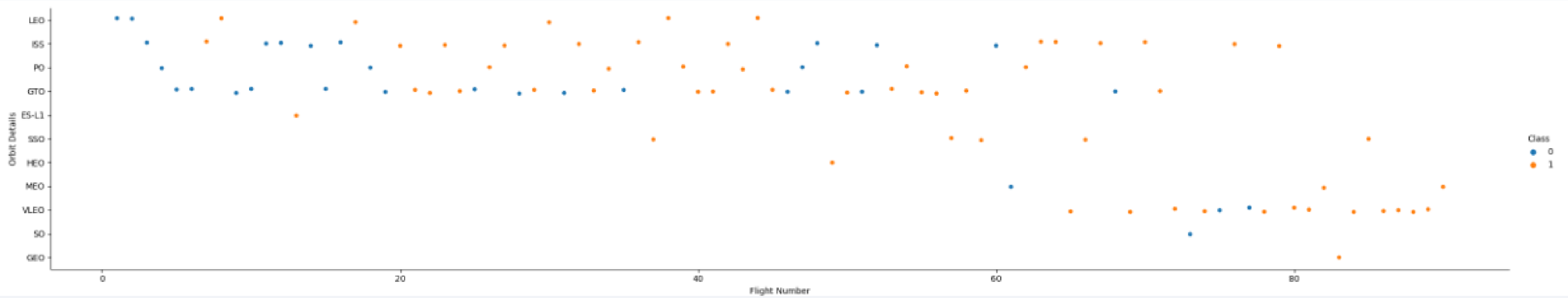
- Key Observations:
  - Across all launch sites, there is a positive correlation between payload mass and success rate. Generally, as the payload mass increases, the likelihood of a successful launch also increases.
  - Specifically, launches with a payload mass over 7000 kg tend to have a higher success rate, indicating that handling heavier payloads is well-managed.
  - Notably, KSC LC 39A stands out with a 100% success rate for payload masses under 5500 kg, demonstrating its exceptional track record in handling smaller payloads effectively.

Success Rate vs. Orbit Type

- Key Observations:
  - There are several orbits that have achieved a 100% success rate, namely ES-L1, GEO, HEO, and SSO. This indicates a high level of reliability and effectiveness in launching missions to these specific orbits.
  - On the other hand, there is an orbit, namely SO, that has experienced a 0% success rate. This suggests challenges or difficulties in achieving successful launches to this particular orbit.
  - There are several orbits with success rates ranging between 50% and 85%. These include GTO (Geostationary Transfer Orbit), ISS (International Space Station), LEO (Low Earth Orbit), MEO (Medium Earth Orbit), and PO (Polar Orbit). This suggests a mixed level of success and room for improvement in achieving consistent and reliable launches to these orbits.

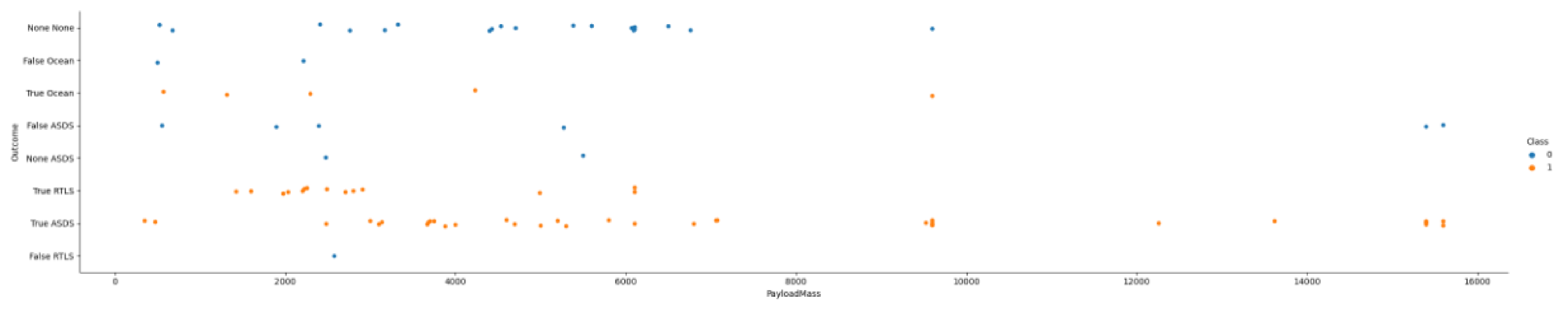


Flight Number vs. Orbit Type



- Key Observations:
  - In the Low Earth Orbit (LEO), there appears to be a correlation between the success rate and the number of flights. As the number of flights in LEO increases, the success rate also tends to increase. This suggests that experience and repetition in launching to LEO contribute to improved success rates.
  - However, in the Geostationary Transfer Orbit (GTO), there seems to be no discernible relationship between the flight number and the success rate. The success rate in GTO does not appear to be influenced by the number of previous flights. Other factors may play a more significant role in determining success in this particular orbit.

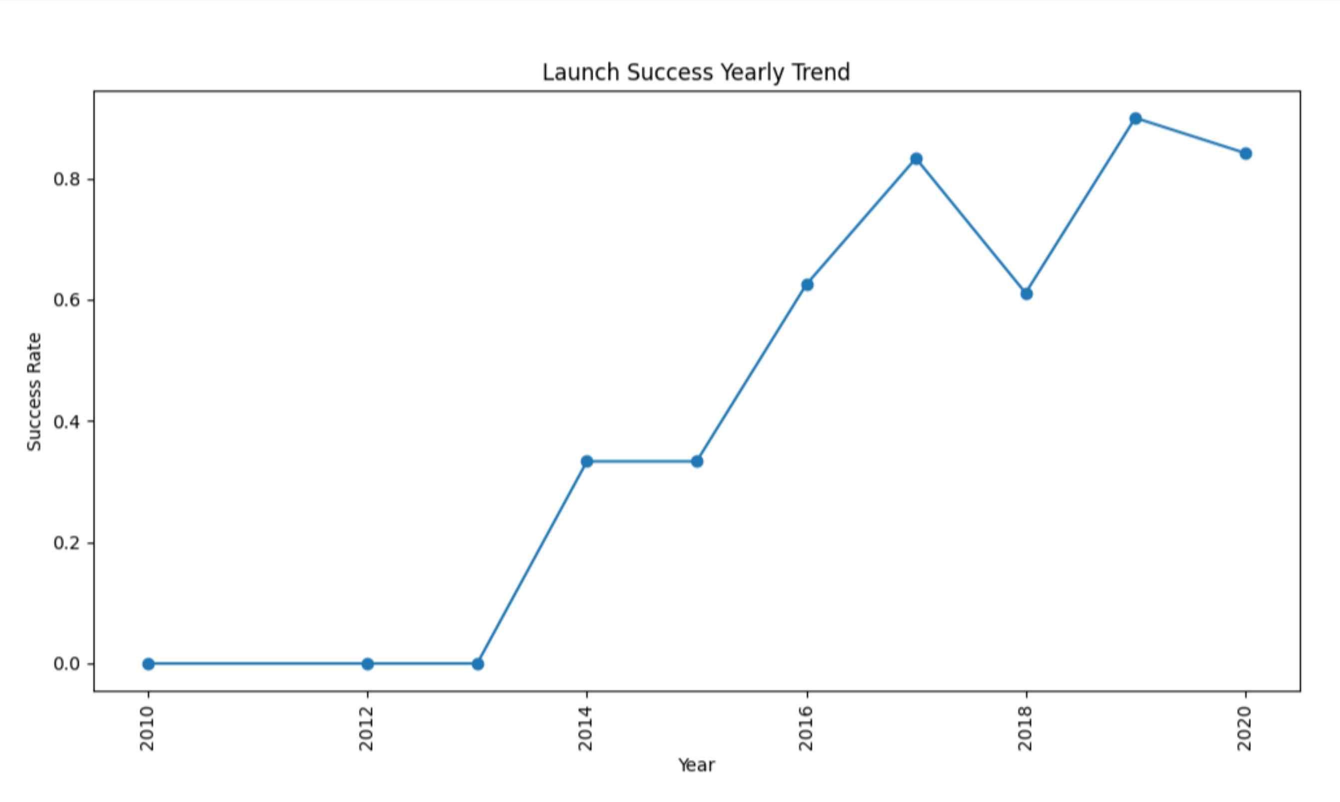
Payload vs. Orbit Type



- Key Observations:
  - Heavy payloads have a negative impact on achieving successful launches in Geostationary Transfer Orbit (GTO) orbits. The higher the payload mass, the lower the success rate in GTO orbits. This suggests that launching heavier payloads to GTO orbits poses challenges or limitations.
  - Conversely, heavy payloads have a positive influence on achieving successful launches in Geostationary Transfer Orbit (GTO) and Polar Low Earth Orbit (LEO) orbits, specifically for the International Space Station (ISS). In these orbits, higher payload masses are associated with higher success rates, indicating effective handling of heavier payloads in these specific mission profiles.

Launch Success Yearly Trend

- Key Observation:
  - The success rate of launches has shown a consistent increase from 2013 to 2020.



All Launch Site Names

Display the names of the unique launch sites in the space mission

In [8]:
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[8]:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A

- Key Observation:
  - The names of the unique launch sites in the space mission are displayed, providing insights into the specific locations from which launches have occurred.

Launch Site Names Begin with 'CCA'

- Key Observation:
  - The dataset includes 5 records of launch sites that begin with the string 'CCA'. These specific launch sites may have common characteristics or operational considerations worth exploring for further analysis.

Display 5 records where launch sites begin with the string 'CCA'

In [9]:

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[9]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

```
n [10]:  sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';

 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
ut[10]:
```

total_payload
111268

- Key Observation:
  - The total payload mass carried by boosters launched by NASA (CRS) provides valuable information about the capacity and capability of NASA's Commercial Resupply Services (CRS) program in delivering payloads to space. Analyzing this data can offer insights into the magnitude and scale of NASA's cargo missions and their contributions to various space exploration endeavors.

Average Payload Mass by F9 v1.1

```
In [11]: sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[11]:

avg_payload
2928

- Key Observation:
  - Analyzing the average payload mass carried by booster version F9 v1.1 allows for an understanding of its performance and suitability for different mission requirements. This information can be useful in assessing the capabilities and efficiency of this particular booster version in delivering payloads to space.

First Successful Ground Landing Date

```
3]:  sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';

 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

3]:

first_success_gp
2015-12-22

- Key Observation:
  - Identifying the date of the first successful landing outcome on a ground pad provides insights into the advancements made in rocket reusability and the technological breakthroughs that have revolutionized space exploration. This achievement has paved the way for subsequent successful landings and has significantly impacted the future of spaceflight.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

[14]:
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Su
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

t[14]:

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

- Key Observation:
  - Analyzing the successful drone ship landings with payload masses between 4000 and 6000 provides valuable insights into the capabilities and reliability of the landing systems. This information indicates that the drone ships have effectively supported the successful recovery of boosters carrying payloads within the specified range, showcasing their importance in enabling missions with medium-sized payloads.

Total Number of Successful and Failure Mission Outcomes

```
In [15]:  sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;

 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[15]:

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

- Key Observation:
  - Analyzing the total number of successful and failed mission outcomes offers valuable insights into the overall performance and success rate of the missions conducted. This information allows for the evaluation of mission effectiveness, identification of potential areas for improvement, and assessment of the overall success and reliability of the space program.

Boosters Carried Maximum Payload

- Key Observation:
  - Identifying the booster versions that have carried the maximum payload mass provides valuable insights into the advancements in rocket technology and engineering. These specific versions have proven to be capable of handling larger payloads, indicating their effectiveness and significance in enabling missions with higher payload requirements.

In [23]:
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[23]:

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

2015 Launch Records

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

```
In [24]: sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR'
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[24]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Key Observation:
  - Analyzing the failed landing outcomes in drone ship for the months in 2015 provides insights into the challenges and difficulties faced during the landing phase of missions. Understanding the specific booster versions and launch sites involved can contribute to identifying potential areas for improvement and optimizing future landing operations.

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

Rank 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

In [25]:

```
sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LAND
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[25]:

landing__outcome	qty
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

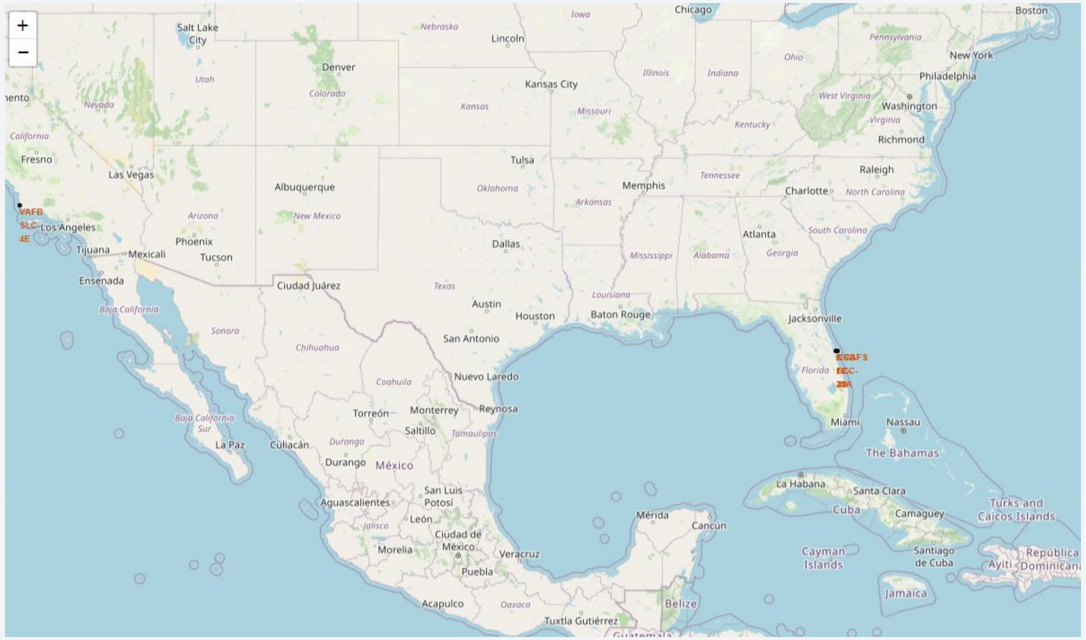
- Key Observation:
  - Ranking the count of landing outcomes in descending order allows for the identification of trends and patterns in the success and failure rates within the specified date range. This information can help assess the overall performance and reliability of the landing systems during that period, highlighting any notable changes or improvements over time.

Section 3

Launch Sites Proximities Analysis

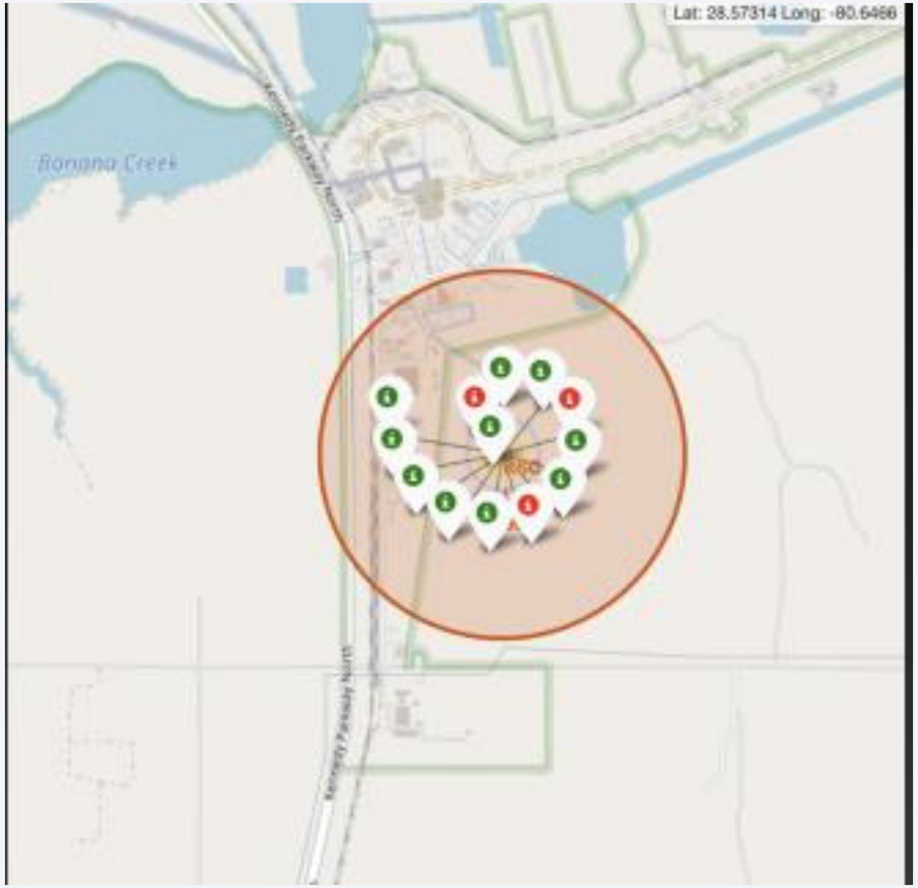
All launch sites' location markers on a global map

- Key Observation:
  - Launching from sites near the Equator allows spacecraft to benefit from the Earth's rotational speed, reducing the energy required for achieving and maintaining orbit. This positioning optimizes launch efficiency and improves the overall success rate of missions.
  - Additionally, all launch sites are situated in close proximity to the coast, enabling rockets to be launched over the ocean. Launching over water minimizes the potential risk of debris or rocket failures affecting populated areas. This safety measure helps protect human lives and property while ensuring the smooth operation of space missions



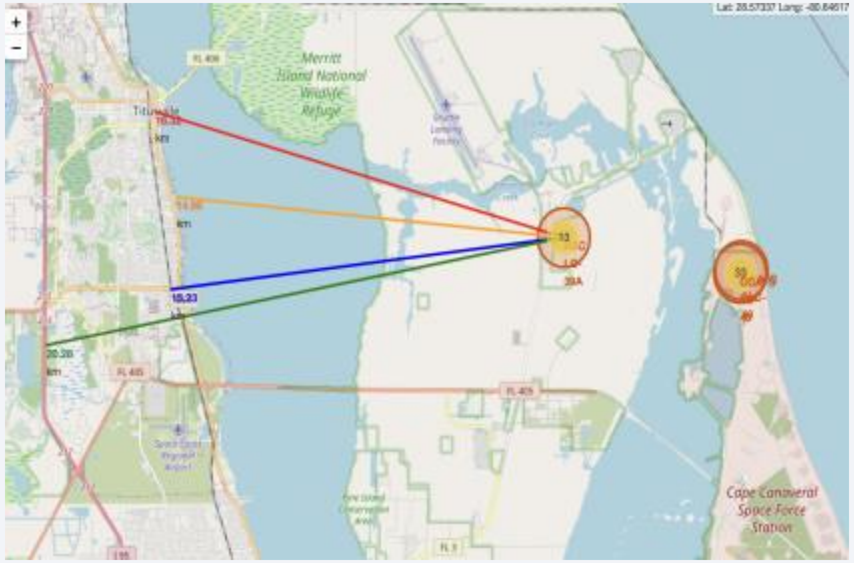
Color-labeled launch records on the map

- Key Observation:
  - Analyzing the distribution of green and red markers allows us to assess the success rates of different launch sites. Sites with a larger number of green markers indicate a higher success rate, while sites with more red markers suggest a lower success rate.
  - Additionally, the specific launch site KSC LC-39A stands out as having a particularly high success rate based on the markers. This observation highlights the effectiveness and reliability of the launch operations carried out at that specific site.



Distance from the launch site KSC LC-39A to its proximities

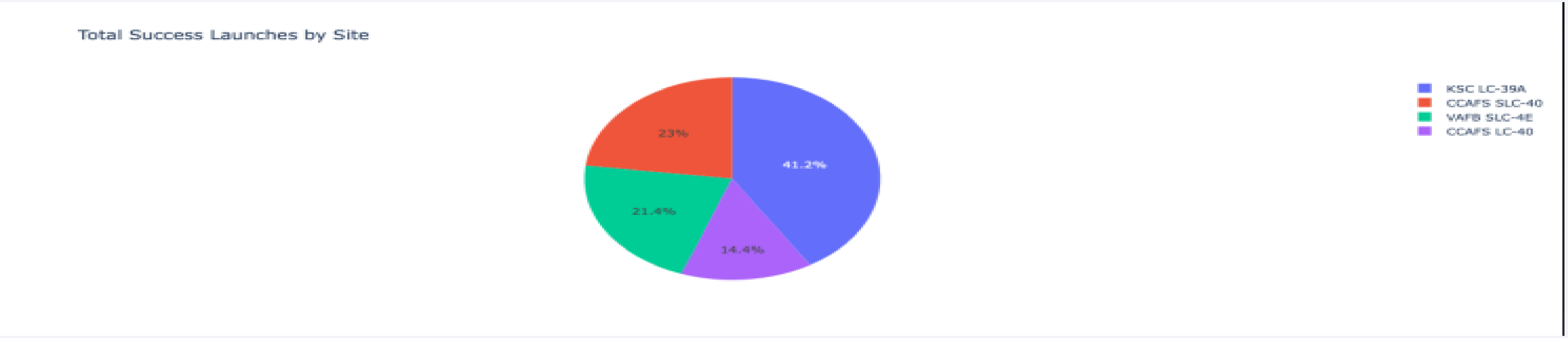
- Key Observation:
  - Assessing the distances between the launch site and nearby features helps evaluate the safety considerations and potential risks associated with rocket launches. Launching from a site that is close to railways, highways, coastlines, and populated areas raises concerns about the safety of those nearby locations in the event of a launch failure.



Section 4

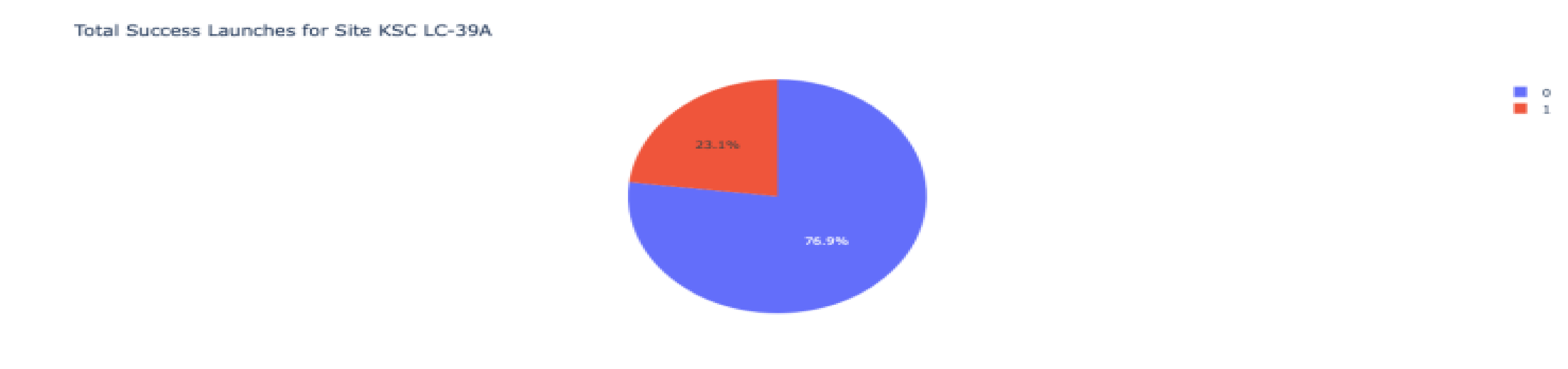
Build a Dashboard with Plotly Dash

Launch success count for all sites



- Key Observation:
  - The dominance of successful launches from KSC LC-39A indicates the site's reliability and effectiveness in conducting space missions. This observation suggests that KSC LC-39A has implemented successful operational procedures, infrastructure, and systems, resulting in a higher success rate for launches carried out from that site.

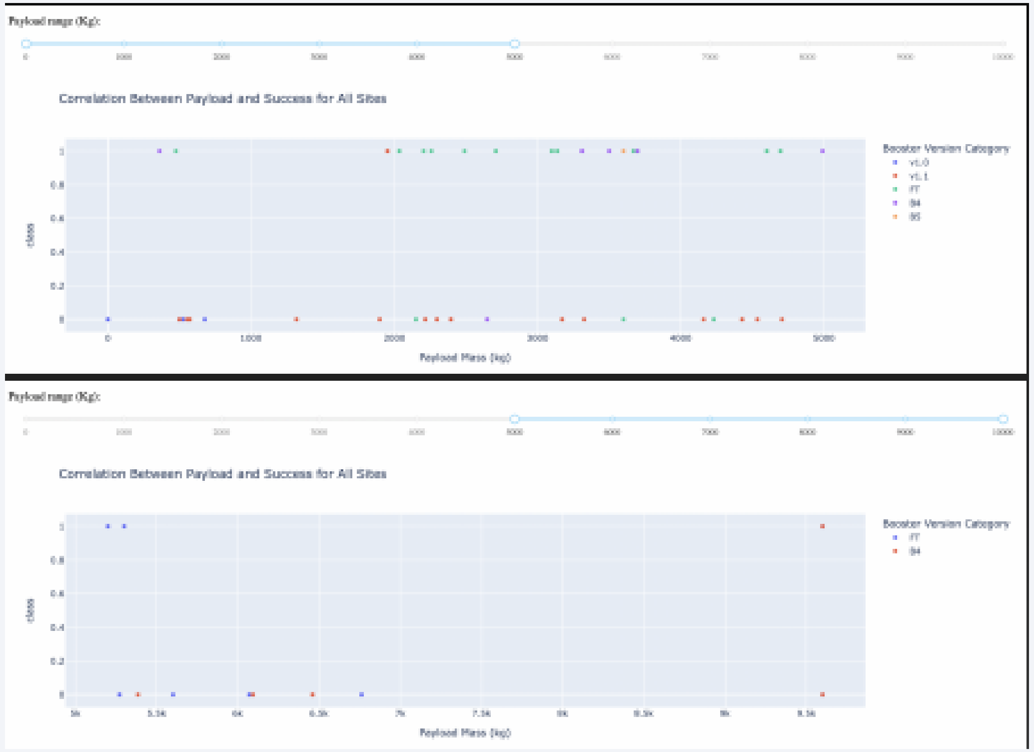
Launch site with highest launch success ratio



- Key Observation:
  - The significant difference in success rates between KSC LC-39A and other sites highlights the reliability and effectiveness of the launch operations carried out at KSC LC-39A. The site's success rate, coupled with a relatively low number of failures, showcases its ability to consistently achieve successful landings and reinforces its reputation as a preferred launch site for space missions.

Payload Mass vs. Launch Outcome for all sites

- Key Observation:
  - Payloads within the 2000-5500 kg range demonstrate a higher likelihood of success, as evidenced by the chart. This suggests that payloads within this specific weight range are well-optimized for successful launches and mission outcomes. It is important for future missions to consider payload mass within this range to maximize the chances of a successful mission.



Section 5

Predictive Analysis (Classification)

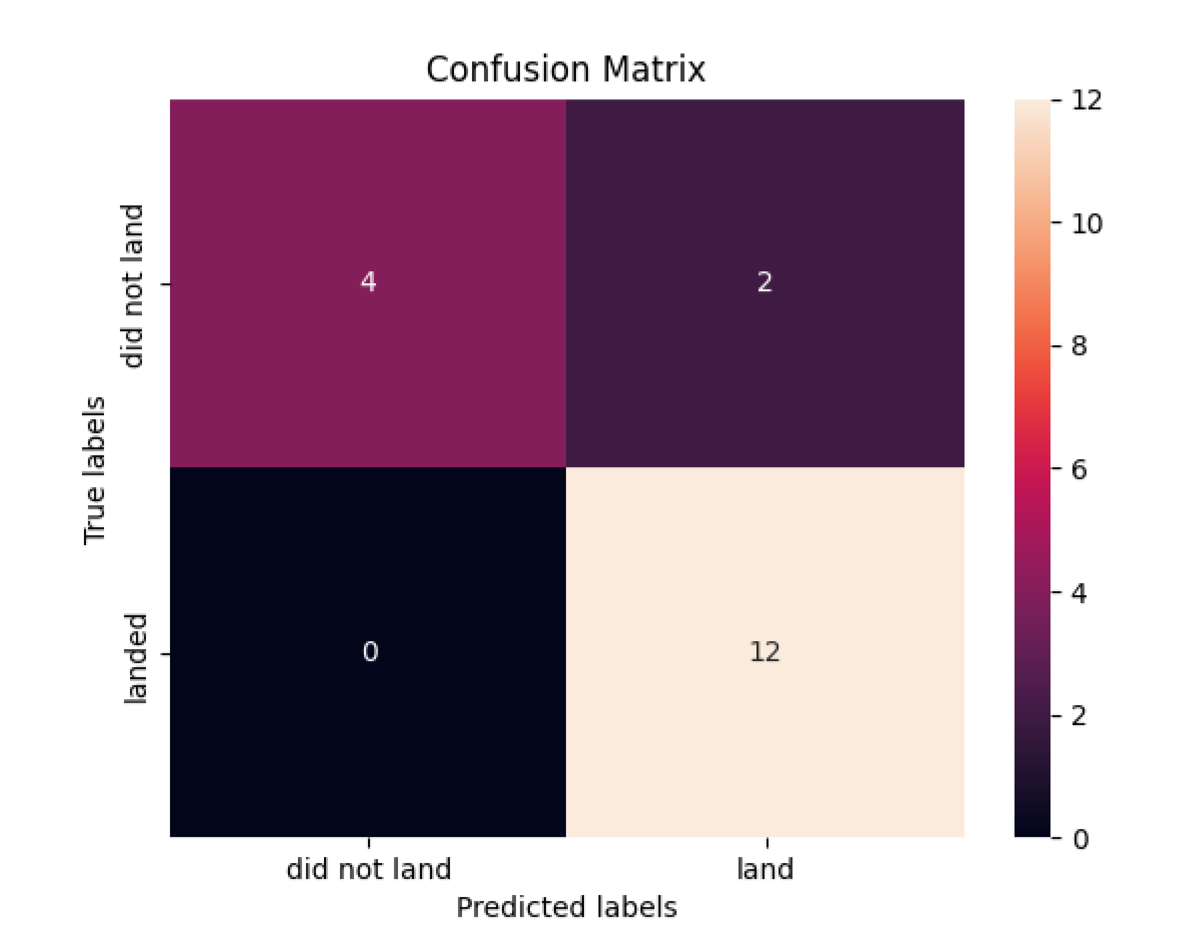
Classification Accuracy

- Key Observation:
  - The Decision Tree Model stands out as the most effective model based on its superior performance and accuracy when evaluated on the complete Dataset.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.90357	0.88889
KNN	0.84821	0.83333

Confusion Matrix

- Key Observation:
  - The presence of false positives indicates that the logistic regression model may have a tendency to predict positive outcomes more frequently than warranted. This observation suggests that the model may need further refinement to reduce false positive rates and improve its overall accuracy.



Conclusions

- The Decision Tree Model is the most effective algorithm for this dataset.
- Launches with lower payload mass show higher success rates.
- Launch sites are typically located near the Equator and in close proximity to the coast.
- The success rate of launches has improved over time.
- KSC LC-39A has the highest success rate among all launch sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.

Appendix

We would like to express our sincere gratitude to the instructors at Coursera, specifically the IBM team, for providing valuable guidance and knowledge throughout the course. Their expertise and support have been instrumental in our learning journey. Thank you for your dedication and commitment to education.

Thank you!