IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Evans B. Sansolis, Ph.D.
September 3, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
  - EDA with Visualization
  - EDA with SQL
  - Interactive Maps with Folium
  - Plotly Dash Dashboard
  - Predictive Analytics
- Conclusion
- Appendix

Executive Summary

Summary of methodologies

- **Data Collection**: Data was collected using SpaceX's REST API and web scraping techniques which includes information about rocket launches, their outcomes, and various related attributes.
- **Data Wrangling**: The collected data was processed to create a success/fail outcome variable which indicates whether a rocket landing was successful or not.
- **Data Exploration**: Data visualization techniques were employed to explore the dataset containing factors such as payload, launch site, flight number, and yearly. Visualization helps in understanding patterns and trends in the data.
- **Data Analysis with SQL:** SQL queries were used to perform various calculations and generate statistics. These calculations may include determining the total payload launched, payload ranges for successful launches, and counting the total number of successful and failed outcomes.
- **Launch Site Success Rates:** The research delved into launch site success rates. This involves analyzing how often launches from different sites resulted in successful landings and considering the proximity of launch sites to geographical markers (which might influence success rates).
- **Visualization of Successful Launches:** Visualization techniques were employed to identify launch sites that had the most success in terms of rocket landings. Successful payload ranges might have been visualized as well.
- **Building Predictive Models:** Machine learning models were developed to predict rocket landing outcomes. The models mentioned include logistic regression, support vector machine (SVM), decision tree, and k-nearest neighbor (KNN). These models use historical data to predict the likelihood of a successful landing based on various input features.

Summary of results

Exploratory Data Analysis:

- Research findings indicate improvements in launch success over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES -L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics:

- Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:

- All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

Project background and context

SpaceX, a prominent player in the space industry, aims to democratize space travel by making it cost-effective for a broader audience. The company has achieved significant milestones, including missions to the International Space Station, the deployment of a satellite network for global internet access, and crewed spaceflights. SpaceX's ability to achieve these feats is largely attributed to the cost-efficiency of its rocket launches, priced at approximately $62 million per launch. This cost advantage arises from SpaceX's innovative practice of reusing the first stage of its Falcon 9 rocket, a strategy that sets it apart from other providers in the industry whose launches can cost upwards of $165 million each.

To determine the pricing of a launch, a key factor is whether the first stage can be successfully recovered and reused. This critical assessment can be made by leveraging publicly available data and employing machine learning models. These models allow for the prediction of whether SpaceX or its competitors can effectively recycle the first stage, ultimately influencing the overall cost of the launch.

Problems you want to find answers

- **Factors Influencing First-Stage Landing Success**: The impact of payload mass, launch site, flight count, and orbital trajectory on the success of first-stage landings.
- **Temporal Evolution of Successful Landings**: The trend in the rate of successful rocket landings as time progresses.
- **Optimal Predictive Model for Landing Success**: Determining the most effective predictive model for classifying successful and unsuccessful rocket landings (binary classification).

Section 1

Methodology

Methodology

Executive Summary

Steps:

Data Collection: Gathered data using the SpaceX REST API and web scraping methods.

Data Wrangling: Prepared the data for analysis and modeling through tasks such as data filtering, handling missing values, and applying one-hot encoding.

Exploratory Data Analysis (EDA): Utilized SQL and data visualization techniques to gain insights from the data.

Data Visualization: Employed Folium and Plotly Dash for visualizing the dataset.

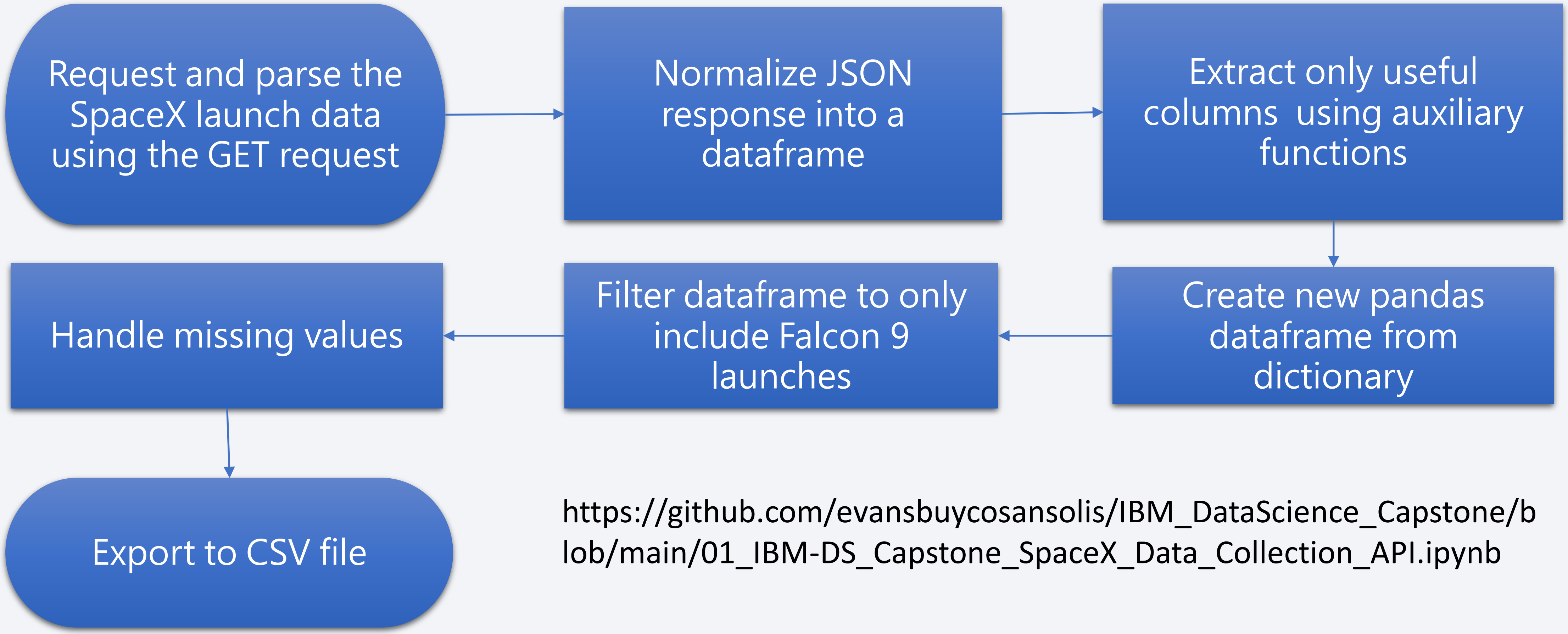
Predictive Modeling: Constructed classification models to predict rocket landing outcomes. This involved model tuning and evaluation to identify the best-performing model and optimal parameters.

Data Collection

Steps:

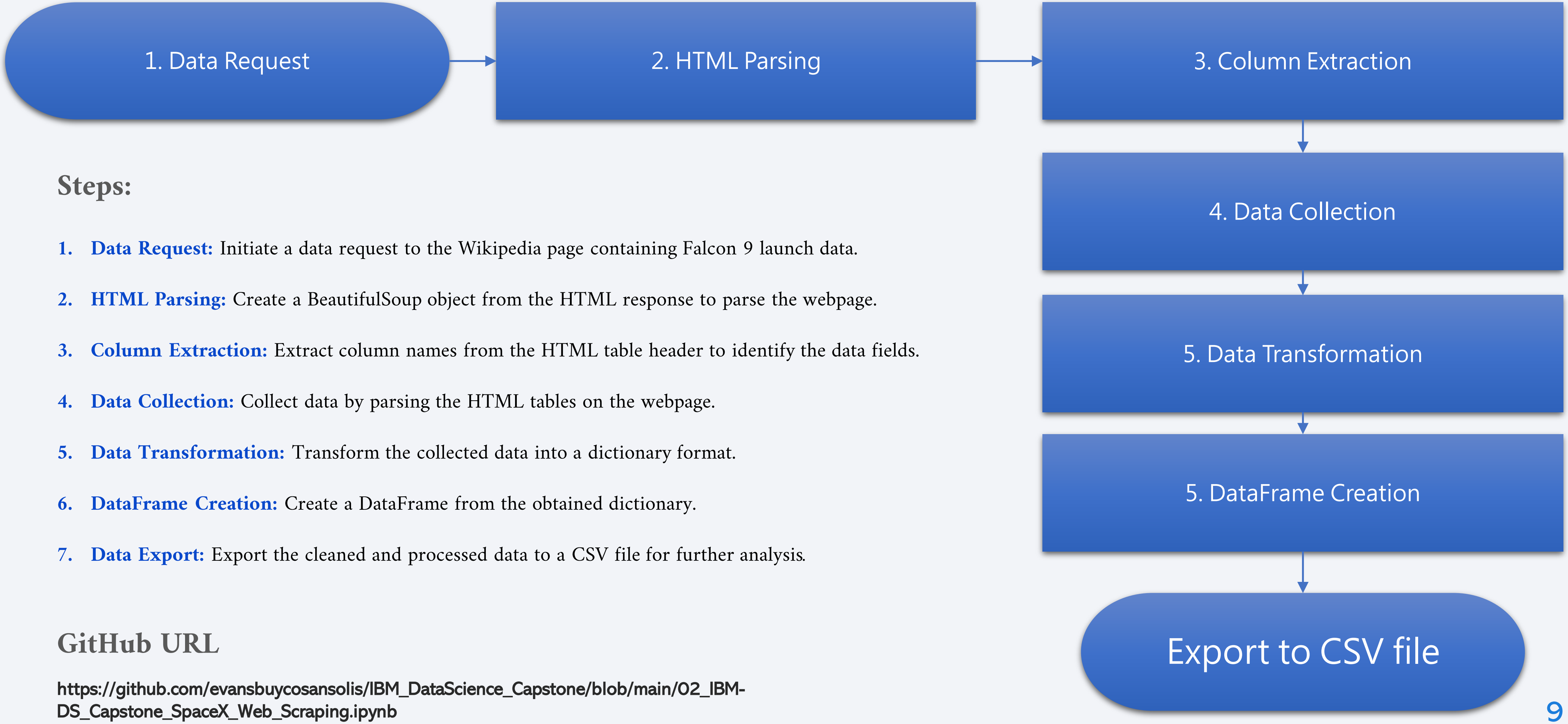
1. **Data Retrieval:** Initiate a request to the SpaceX API to obtain rocket launch data.
2. **Data Decoding:** Decode the API response using the .json() method and convert it into a DataFrame using .json_normalize().
3. **Custom Functions:** Use custom functions to request specific launch information from the SpaceX API.
4. **Data Transformation:** Transform the retrieved data into a dictionary format.
5. **DataFrame Creation:** Create a DataFrame from the obtained dictionary.
6. **Data Filtering:** Filter the DataFrame to include only Falcon 9 launches.
7. **Handling Missing Values:** Replace missing values in the Payload Mass column with the calculated mean value.
8. **Data Export:** Export the cleaned and processed data to a CSV file for further analysis.

Data Collection – SpaceX API



https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/blob/main/01_IBM-DS_Capstone_SpaceX_Data_Collection_API.ipynb

Data Scraping



Steps:

1. **Data Request:** Initiate a data request to the Wikipedia page containing Falcon 9 launch data.
2. **HTML Parsing:** Create a BeautifulSoup object from the HTML response to parse the webpage.
3. **Column Extraction:** Extract column names from the HTML table header to identify the data fields.
4. **Data Collection:** Collect data by parsing the HTML tables on the webpage.
5. **Data Transformation:** Transform the collected data into a dictionary format.
6. **DataFrame Creation:** Create a DataFrame from the obtained dictionary.
7. **Data Export:** Export the cleaned and processed data to a CSV file for further analysis.

GitHub URL

https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/blob/main/02_IBM-DS_Capstone_SpaceX_Web_Scraping.ipynb

Data Collection Wrangling

Calculate the number of launches on each site → Calculate the number and occurrence of each orbit → Calculate the number and occurrence of mission outcome per orbit type → Create a landing outcome label from Outcome column using one-hot encoding → Export to CSV file

Steps:

Data Preprocessing Steps:

1. Perform Exploratory Data Analysis (EDA) to understand the dataset and determine data labels.
2. Calculate the number of launches for each launch site.
3. Calculate the number and occurrence of different orbit types.
4. Calculate the number and occurrence of mission outcomes per orbit type.

Creating a Binary Landing Outcome Column (dependent variable) based on the provided criteria:

- Landing was not always successful (0)
- True Ocean: mission outcome had a successful landing to a specific region of the ocean (1)
- False Ocean: represented an unsuccessful landing to a specific region of the ocean (0)
- True RTLS: meant the mission had a successful landing on a ground pad (1)
- False RTLS: represented an unsuccessful landing on a ground pad (0)
- True ASDS: meant the mission outcome had a successful landing on a drone ship (1)
- False ASDS: represented an unsuccessful landing on a drone ship (0)

Data Export:

- Export the preprocessed data, including the binary landing outcome column, to a CSV file for further analysis and modeling.

GitHub URL

https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/blob/main/03_IBM-DS_Capstone_SpaceX_Data_Wrangling.ipynb

EDA with SQL

Display Queries:

1. ***Names of unique launch sites***: This query retrieves and displays the names of all unique launch sites.
2. *5 **records where the launch site begins with 'CCA'***: This query retrieves and displays 5 records of launch sites where the name starts with 'CCA'.
3. ***Total payload mass carried by boosters launched by NASA (CRS)***: This query calculates and displays the total payload mass carried by boosters launched by NASA under the Commercial Resupply Services (CRS) program.
4. ***Average payload mass carried by booster version F9 v1.1***: This query calculates and displays the average payload mass carried by booster versions with the name 'F9 v1.1'.

List Queries:

1. ***Date of the first successful landing on a ground pad***: This query retrieves and displays the date of the first successful landing of a booster on a ground pad.
2. ***Names of boosters that had a successful landing on a drone ship and had a payload mass greater than** 4,000 **but less than** 6,000*: This query lists the names of boosters that had successful landings on a drone ship and carried a payload mass within the specified range.
3. ***Total number of successful and failed missions***: This query calculates and displays the total number of both successful and failed missions.
4. ***Names of booster versions that have carried the max payload***: This query retrieves and lists the names of booster versions that have carried the maximum payload mass.
5. ***Failed landing outcomes on a drone ship, their booster version, and launch site for the months in the year** 2015*: This query lists the failed landing outcomes on a drone ship, along with their booster version and launch site, specifically for the months in the year 2015.
6. ***Count of landing outcomes between** 2010-06-04 **and** 2017-03-20 **(descending order)***: This query counts and displays the landing outcomes within the specified date range, sorted in descending order of count.

GitHub URL

https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/blob/main/04_IBM-DS_Capstone_SpaceX_EDA_SQL.ipynb

EDA with Data Visualization

Charts:

1. **Flight Number vs. Payload:** This chart displays the relationship between the flight number and the payload of the rocket launches.
2. **Flight Number vs. Launch Site:** This chart illustrates how the flight number is related to the launch site, providing insights into launch site usage over time.
3. **Payload Mass (kg) vs. Launch Site:** This chart shows how the payload mass varies based on the launch site, helping us understand if certain launch sites are preferred for heavier payloads.
4. **Payload Mass (kg) vs. Orbit Type:** This chart visualizes the relationship between payload mass and the type of orbit, allowing us to see if there's a correlation between payload mass and orbit choice.

EDA with Visualization Analysis:

1. **View relationship by using scatter plots:** Scatter plots are used to explore the relationship between two variables. If a discernible pattern or trend exists in the scatter plot, it suggests a potential relationship that may be useful for machine learning.

2. **Show comparisons among discrete categories with bar charts:** Bar charts are useful for comparing different categories or groups and understanding how they relate to a measured value. This helps in identifying patterns and differences within categorical data.

GitHub URL

https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/blob/main/05_IBM-DS_Capstone_SpaceX_EDA_Data_Visualization.ipynb

Build an Interactive Map with Folium

Markers Indicating Launch Sites:

- A blue circle was added at NASA Johnson Space Center's coordinates on the map.
- A popup label displaying the name of NASA Johnson Space Center was included using its latitude and longitude coordinates.
- Red circles were added at the coordinates of all launch sites.
- Popup labels showing the names of each launch site were included using their respective latitude and longitude coordinates.

Colored Markers of Launch Outcomes:

- Colored markers were added to represent successful launches (green) and unsuccessful launches (red) at each launch site.
- This visual representation allows for the identification of launch sites with high success rates (green) and those with lower success rates (red).

Distances Between a Launch Site to Proximities:

- Colored lines were added to the map to illustrate the distances between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city.
- This provides a visual reference to understand the proximity of the launch site to these important geographical features.

GitHub URL

https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/blob/main/06_IBM-DS_Capstone_SpaceX_Interactive_Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites:

- Users can choose to view all launch sites or select a specific launch site from the dropdown list.

Slider of Payload Mass Range:

- Users can adjust the payload mass range using a slider, allowing them to filter launches within a specific payload mass range.

Pie Chart Showing Successful Launches:

- A pie chart displays the percentage of successful and unsuccessful launches as part of the total launches. Users can visualize the success rate briefly.

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version:

- Users can view a scatter chart that illustrates the correlation between payload mass and launch success.

- The chart also allows users to explore how different booster versions perform in terms of payload mass and success rate.

GitHub URL

https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/blob/main/07_IBM-DS_Capstone_SpaceX_Interactive_Visual_Analytics_Plotly.py

Predictive Analysis (Classification)

Data Preparation:

- Create a NumPy array from the Class column.
- Standardize the data using StandardScaler. Fit and transform the data.

Data Splitting:

- Split the data into training and testing sets using train_test_split.

Model Selection:

- Create a GridSearchCV object with cv=10 for parameter optimization. This step is important for finding the best hyperparameters for each algorithm.

Model Training and Evaluation:

- Apply GridSearchCV on different machine learning algorithms, including logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), and K-Nearest Neighbor (KNeighborsClassifier()).
- Calculate accuracy on the test data using the .score() method for all models.
- Assess the confusion matrix for all models to understand their performance in terms of true positives, true negatives, false positives, and false negatives.

Model Comparison:

- Identify the best model using evaluation metrics such as Jaccard Score, F1 Score, and Accuracy. These metrics will help determine which model performs the best for the given data.

GitHub URL

https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/blob/main/08_IBM-DS_Capstone_SpaceX_Predictive_Analytics.ipynb

Results Summary

Exploratory Data Analysis:

- The analysis shows that launch success has improved over time, indicating advancements in space technology and processes.
- Among the different landing sites, KSC LC-39A has the highest success rate, making it a preferred choice for launches.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate, suggesting these orbits are well-understood and suitable for successful missions.

Visual Analytics:

- Most launch sites are located near the equator, which is a common practice in space launches due to the rotational speed advantage.
- Launch sites are strategically positioned to be far enough away from populated areas (city), and transportation routes (highway, railway), while still being accessible for logistical support.

Predictive Analytics:

- The Decision Tree model is identified as the best predictive model for the dataset. This model likely provides the most accurate predictions of landing outcomes based on the selected features and data.

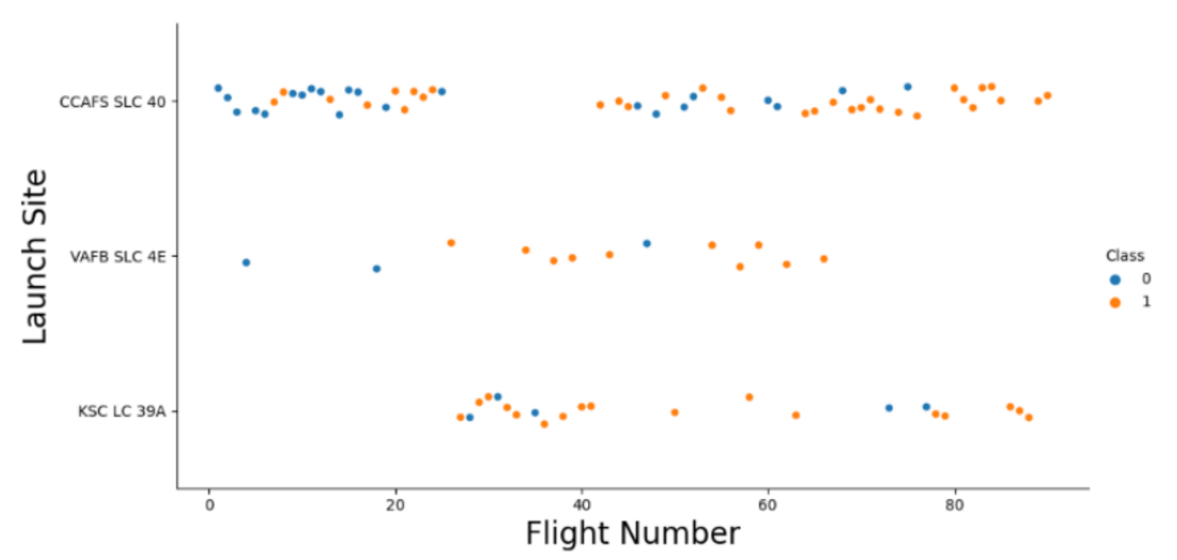
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis:

- The success rate of earlier flights (represented by blue) was lower compared to later flights (represented by orange), suggesting an improvement in launch success over time. This trend aligns with the previous observation that launch success has improved over time.
- Approximately half of the launches originated from the CCAFS SLC 40 launch site, indicating its significant usage.
- Launch sites such as VAFB SLC 4E and KSC LC 39A have higher success rates, implying that these sites may have favorable conditions or procedures for successful launches.
- The inference that new launches have a higher success rate aligns with the overall trend of improved success over time. This observation suggests that advancements in technology or experience have contributed to the higher success rates in more recent launches.



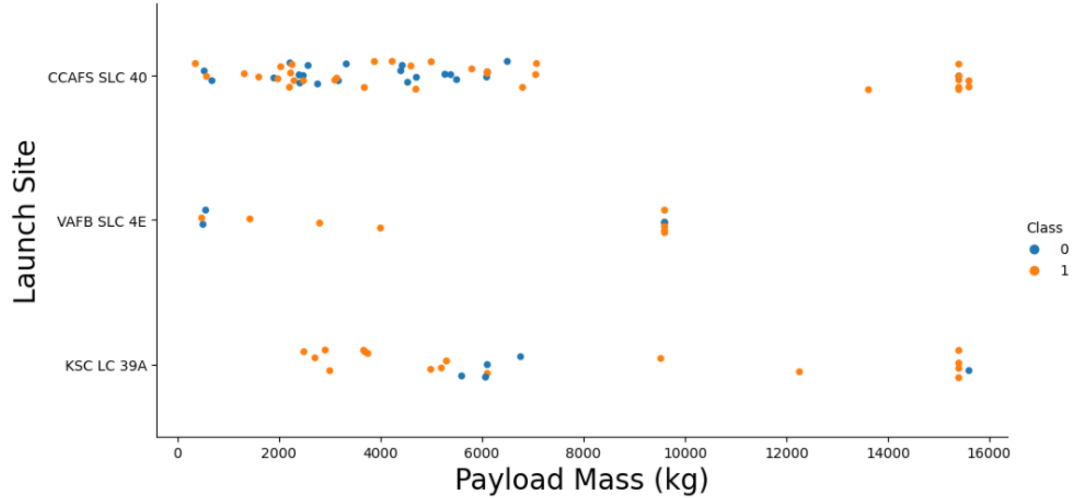
These additional findings enhance the understanding of launch success trends and factors influencing outcomes in the dataset.

Payload vs. Launch Site

Exploratory Data Analysis:

- The finding that higher payload mass is generally associated with a higher success rate suggests that launches with heavier payloads tend to have a better chance of success. This relationship between payload mass and success can be an important factor for mission planning.
- The observation that most launches with a payload greater than 7,000 kg were successful highlights a threshold where payload mass becomes a critical factor in launch success. This threshold can be an important consideration when determining the feasibility of launching heavier payloads.
- The 100% success rate for launches less than 5,500 kg at KSC LC 39A indicates that this launch site is particularly well-suited for smaller payloads. This information can be valuable for selecting the appropriate launch site based on payload size.
- The observation that VAFB SLC 4E has not launched anything greater than approximately 10,000 kg suggests that this launch site may have limitations on handling very heavy payloads. Understanding such limitations can inform decisions on launch site selection and payload planning.

In [40]:
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df,aspect=2)
plt.xlabel("Payload Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



These insights contribute to a deeper understanding of the relationship between payload characteristics, launch success, and the suitability of different launch sites for various payload sizes..

Success Rate vs. Orbit Type

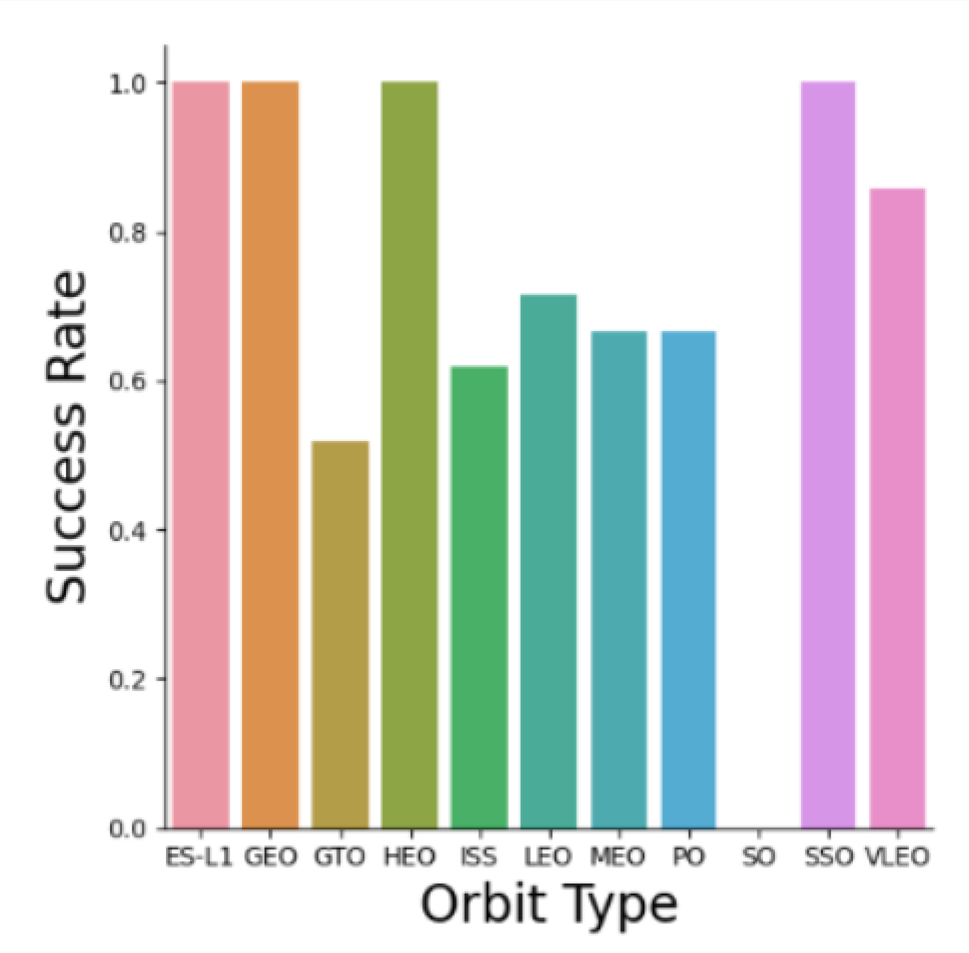
Exploratory Data Analysis:

- **100% Success Rate (ES-L1, GEO, HEO, and SSO):** These orbital destinations have consistently achieved successful launches. This indicates a high level of reliability for missions targeting these orbits, which can be crucial for satellite deployments and other space operations.

- **50%-80% Success Rate (GTO, ISS, LEO, MEO, PO):** These orbital destinations have success rates ranging from 50% to 80%. This suggests that missions to these orbits have had a moderate level of success, with some variability. Further analysis can help identify factors contributing to both successful and unsuccessful launches in these categories.

- **0% Success Rate (SO):** The 0% success rate for the SO (Sub-Orbital) category indicates that all missions to this type of destination have been unsuccessful. Investigating the reasons for these failures can provide valuable insights into the challenges associated with sub-orbital missions.

These success rate categories offer a useful framework for evaluating the performance of launches to different orbital destinations. Further analysis can delve into the specific factors, such as launch sites, payload characteristics, or mission parameters, that may contribute to the observed success rates in each category.



Flight Number vs. Orbit Type

Exploratory Data Analysis:

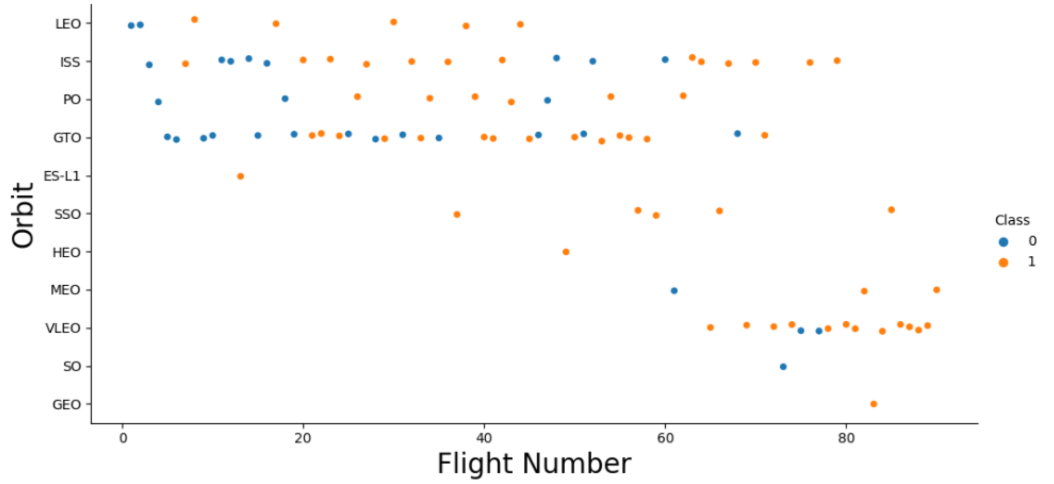
- **Success Rate Increases with the Number of Flights:** In general, there appears to be a positive correlation between the number of flights to a particular orbit and the success rate. This suggests that as more missions are conducted to a specific orbit, operators and engineers gain experience and expertise, leading to improved success rates.

- **LEO Orbit:** The LEO (Low Earth Orbit) stands out as an example where this relationship is highly apparent. As the number of LEO missions increases, the success rate also tends to increase. This could be attributed to the fact that LEO missions are relatively common and often serve purposes like satellite deployments, where operators have had more opportunities to learn and refine their processes.

- **GTO Orbit:** Interestingly, the GTO (Geostationary Transfer Orbit) does not follow the same trend. Even though there are more flights to the GTO orbit, the success rate does not show a consistent increase. This may indicate that GTO missions are more complex or have unique challenges that do not necessarily improve with the number of flights.

Further analysis could explore the specific factors contributing to these trends, such as mission complexity, launch site variations, or changes in technology over time. Understanding these dynamics can provide valuable insights into the factors influencing mission success for different orbital destinations.

In [44]:
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df,aspect=2)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Payload vs. Orbit Type

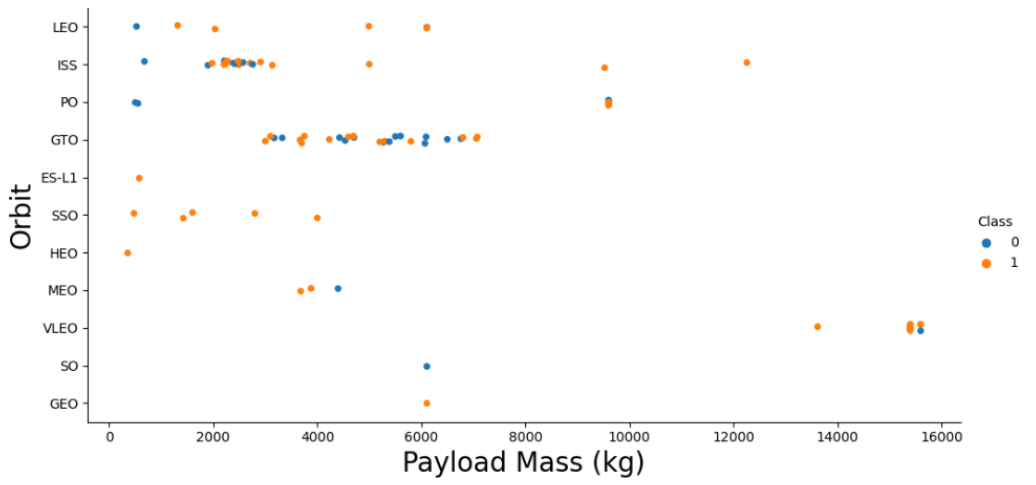
Exploratory Data Analysis:

Payload Mass and Orbital Destinations: The success rate of missions appears to be influenced by the payload mass and the specific orbital destination. Here are some key takeaways:.

- **LEO (Low Earth Orbit), ISS (International Space Station), and PO (Polar Orbit):** These orbital destinations seem to have a positive correlation between heavy payloads and success rate. In other words, missions with heavier payloads tend to have higher success rates in these orbits. This may be because these orbits often involve missions like satellite deployments, where heavier payloads indicate larger and potentially more complex satellites. The higher success rates for heavy payloads in these orbits could be attributed to the maturity of technology and experience in launching such missions.

- **GTO (Geostationary Transfer Orbit):** The GTO orbit exhibits a more mixed relationship between payload mass and success rate. This suggests that GTO missions are more complex, and success is influenced by various factors beyond payload mass alone. Some GTO missions with heavy payloads may succeed, while others may not, depending on mission-specific challenges.

In [46]:
```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df,aspect=2)
plt.xlabel("Payload Mass (kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Understanding these relationships between payload mass and orbital destinations can help space agencies and organizations make more informed decisions when planning and executing missions. It highlights the importance of considering the specific requirements and challenges associated with different orbits and payload masses to maximize mission success.

Launch Success Yearly Trend

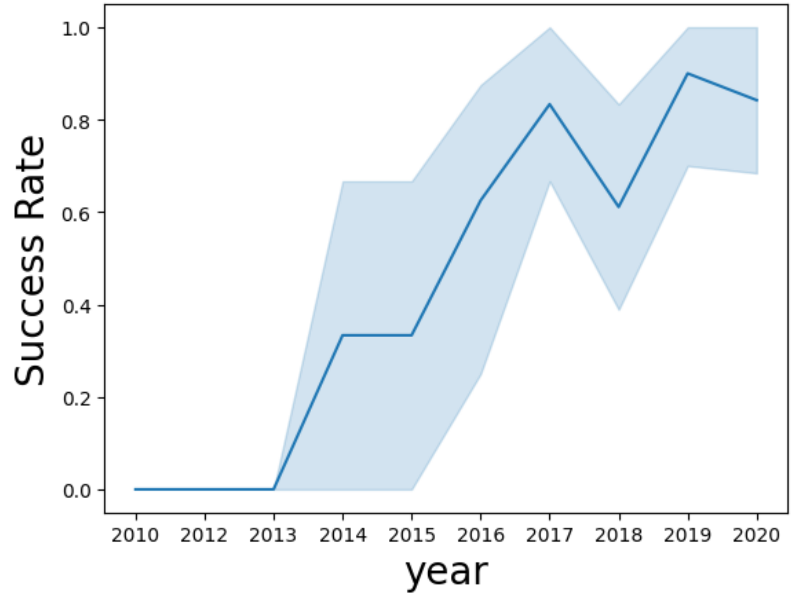
Exploratory Data Analysis:

Temporal Trends in Launch Success Rates: Your analysis reveals several key trends in launch success rates over time:.

- **Improvement from 2013-2017:** The success rate of space launches showed a positive trend from 2013 to 2017. This period of improvement suggests that space agencies and organizations may have made advancements in technology, procedures, and mission planning that contributed to higher success rates during these years.
- **Decrease from 2017-2018:** There was a decrease in launch success rates from 2017 to 2018. This could be due to various factors, such as the complexity of missions attempted during that period, technical challenges, or specific mission failures. Further analysis could help identify the specific reasons behind this decrease.
- **Improvement from 2018-2019:** Launch success rates improved again from 2018 to 2019. This recovery indicates that any challenges or issues that might have led to the decrease in 2017-2018 were potentially addressed, leading to better performance in subsequent years.
- **Decrease from 2019-2020:** Another decrease in success rates occurred from 2019 to 2020. As with the earlier decrease, understanding the reasons behind this decline would require more in-depth analysis. Factors such as mission complexity, technical issues, or external factors could have contributed.
- **Overall Improvement Since 2013:** Despite some fluctuations, your analysis indicates an overall improvement in launch success rates since 2013. This suggests that the space industry has made progress in enhancing the reliability and success of space missions over the years.

These temporal trends are valuable for space agencies, organizations, and researchers as they provide insights into historical performance and can inform future mission planning and improvements. Identifying the specific factors contributing to success or failure during these periods could lead to further advancements in space exploration and technology.

```
In [49]: # Plot a line chart with x axis to be the extracted year and y axis to be the success rate
         sns.lineplot(data=df, x="Date", y="Class")
         plt.xlabel("year",fontsize=20)
         plt.ylabel("Success Rate",fontsize=20)
         plt.show()
```



All Launch Site Names

[30]:	launch_site
	CCAFS LC-40
	CCAFS SLC-40
	KSC LC-39A
	VAFB SLC-4E

Launch Sites Names:

- CCAFS LC-40: Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40: Cape Canaveral Air Force Station Space Launch Complex 40
- KSC LC-39A: Kennedy Space Center Launch Complex 39A
- VAFB SLC-4E: Vandenberg Air Force Base Space Launch Complex 4E.

These launch sites play a crucial role in space missions and are where rockets are prepared, fueled, and launched into space. Each site may have unique capabilities and is used for specific types of missions and launch vehicles.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- The LIMIT and LIKE clauses were used to display only the top five results where the launch_site name starts with 'CCA'.

```
In [26]: %sql SELECT * \
             FROM SPACEXTBL \
             WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;

 * sqlite:///my_data1.db
Done.

Out[26]:
```

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

Total Payload Mass

- Total payload carried by boosters from NASA
- The SUM() function was used to calculate the total payload carried by boosters from NASA from the payload_mass__kg column.

In [27]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

Out[27]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
- The AVG() function was used to calculate the average payload the average payload mass carried by booster version F9 v1.1
- The WHERE clause was used to filter results so that the calculations were only performed on booster_versions only if named“F9 v1.1

```
In [28]: %sql SELECT AVG(PAYLOAD_MASS__KG_) \
             FROM SPACEXTBL \
             WHERE BOOSTER_VERSION = 'F9 v1.1';

 * sqlite:///my_data1.db
Done.

Out[28]:
```

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Dates of the first successful landing outcome on a ground pad
- The MIN(DATE) function was used to find the date of the first successful landing outcome on a ground pad
- The WHERE clause ensured that the results were filtered to match only when the 'landing_outcome'column is 'Success (ground pad)'

```
In [38]: %sql SELECT MIN(DATE) \
         FROM SPACEXTBL \
         WHERE Landing_Outcome = 'Success (ground pad)'

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

Out[38]:

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters that have successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000
- The BETWEEN clause was used to retrieve the results of payload mass greater than 4000 but less than 6000. The WHERE clause filtered the results to include only boosters that successfully landed on a drone ship.

```
In [39]: %sql SELECT PAYLOAD \
         FROM SPACEXTBL \
         WHERE Landing_Outcome = 'Success (drone ship)' \
         AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

 * sqlite:///my_data1.db
Done.

Out[39]:
```

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failed mission outcomes
- The COUNT() function is used to count the number of occurrences of different mission outcomes with the help of the GROUPBY clause applied to the 'mission_outcome' column. A list of the total number of successful and failed mission outcomes is returned.
- There have been 99 successful mission outcomes out of 101 missions.

```
In [31]: %sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

Out[31]:

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

Boosters Carried Maximum Payload

- Names of the boosters which have carried the maximum payload mass
- The MAX() function was used in a subquery to retrieve a list of boosters that have carried the maximum payload mass.

In [32]:
```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[32]:

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

2015 Launch Records

- The failed landing_outcomes in drone ships, their booster versions, and launch site names for in year 2015
- 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

```
In [47]: te, 6, 2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome like 'Failure%dron%ship%';

 * sqlite:///my_data1.db
Done.

Out[47]:
```

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
01	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
04	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
06	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

- Count ranking of the landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order
- The COUNT() function was used to count the different landing outcomes. The WHERE and BETWEEN clauses filtered the results to only include results between 2010-06-04 and 2017-03-20. The GROUPBY clause ensures that the counts were grouped by their outcome. The ORDERBY and DESC clauses were used to sort the results by descending order.

```
In [48]: ng_Outcome from SPACEXTBL  where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Landing_Count DESC;

 * sqlite:///my_data1.db
Done.

Out[48]:
```

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

Section 3

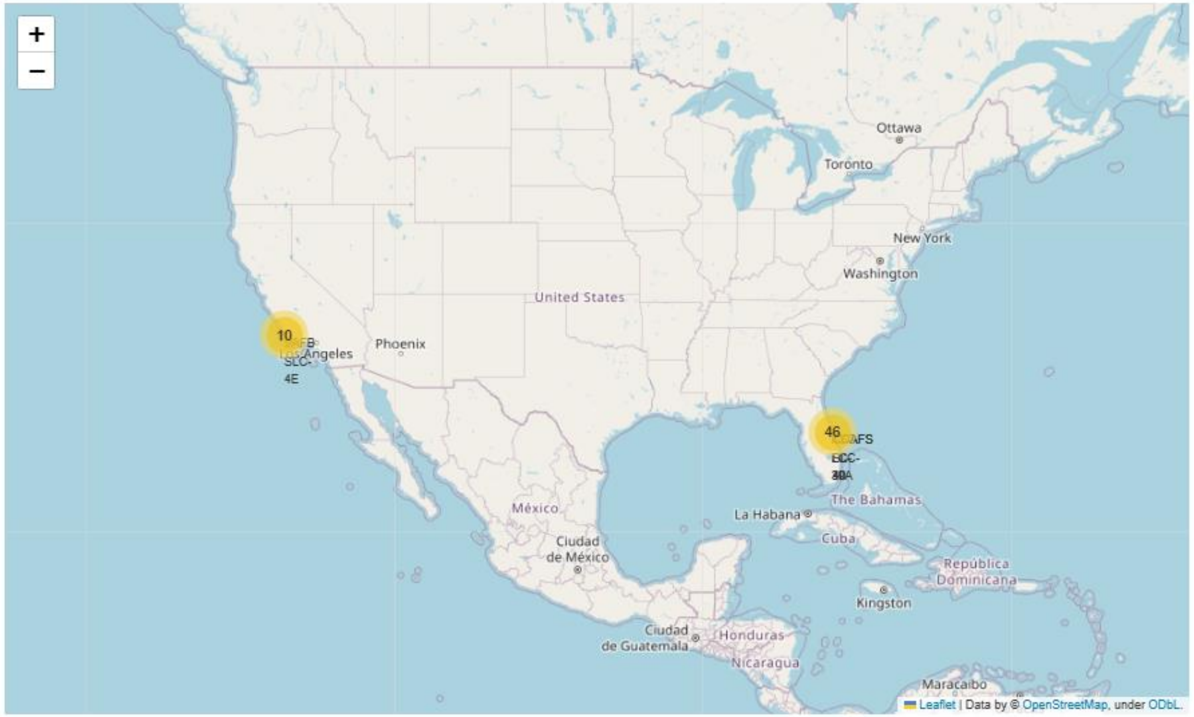
Launch Sites Proximities Analysis

SpaceX Launch Sites Locations

In [37]:
```python
# Add marker_cluster to current site_map
site_map.add_child(marker_cluster)

# for each row in spacex_df data frame
# create a Marker object with its coordinate
# and customize the Marker's icon property to indicate if this launch was successed or failed,
# e.g., icon=folium.Icon(color='white', icon_color=row['marker_color']
for site_lat, site_long, marker_color in zip(spacex_df['Lat'], spacex_df['Long'], spacex_df['marker_color']):
    site_coordinate = [site_lat, site_long]
    marker = folium.map.Marker(
        site_coordinate,
        # Create an icon as a text label
        icon=folium.Icon(color='white',
                         icon_color=marker_color)
    )
    marker.add_to(marker_cluster)
site_map
```

Out[37]:

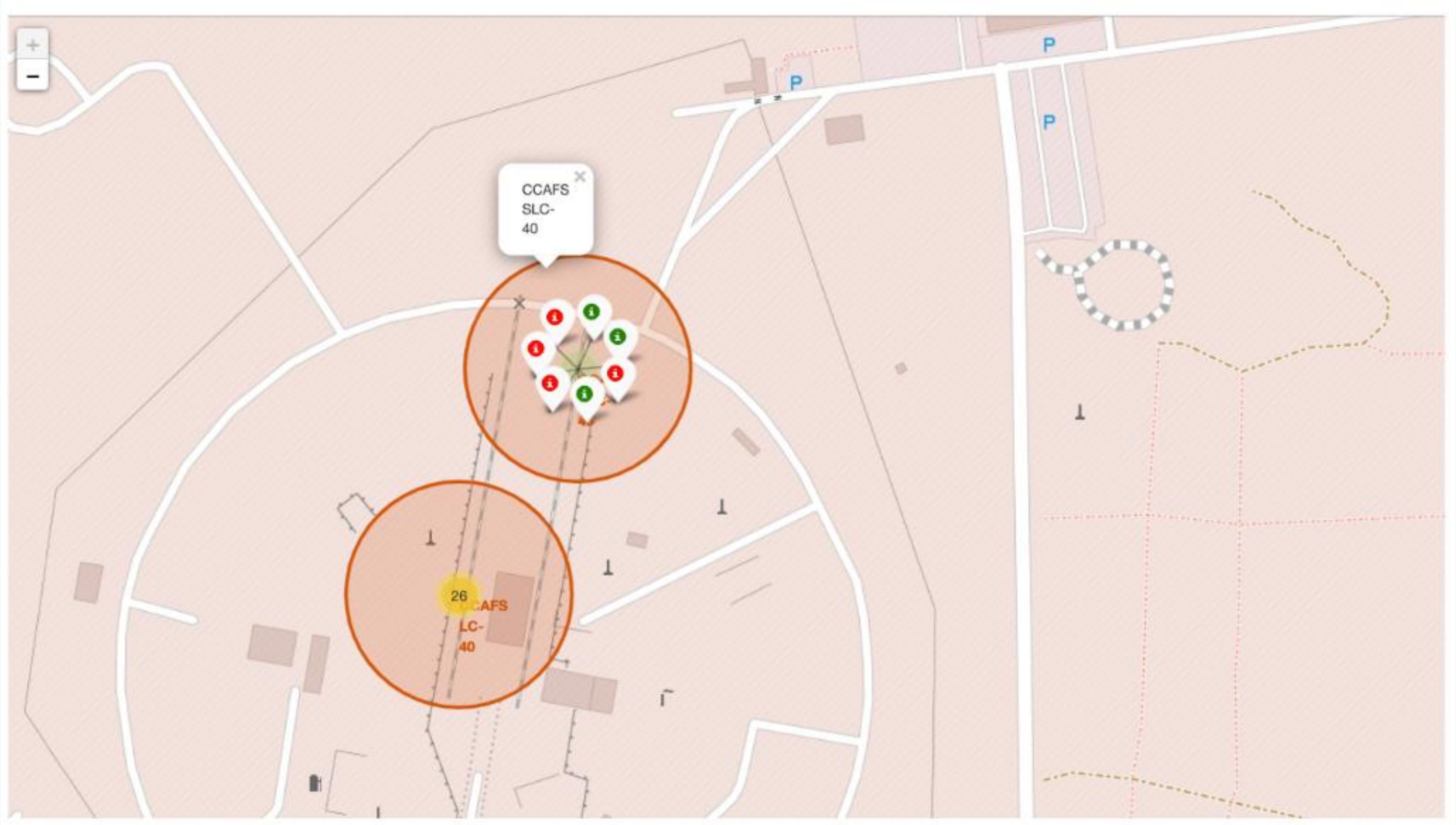


- The yellow markers are indicators of where the locations of all the SpaceX launch sites are situated in the US.
- The launch sites have been strategically placed near the coast.

Launch Outcome

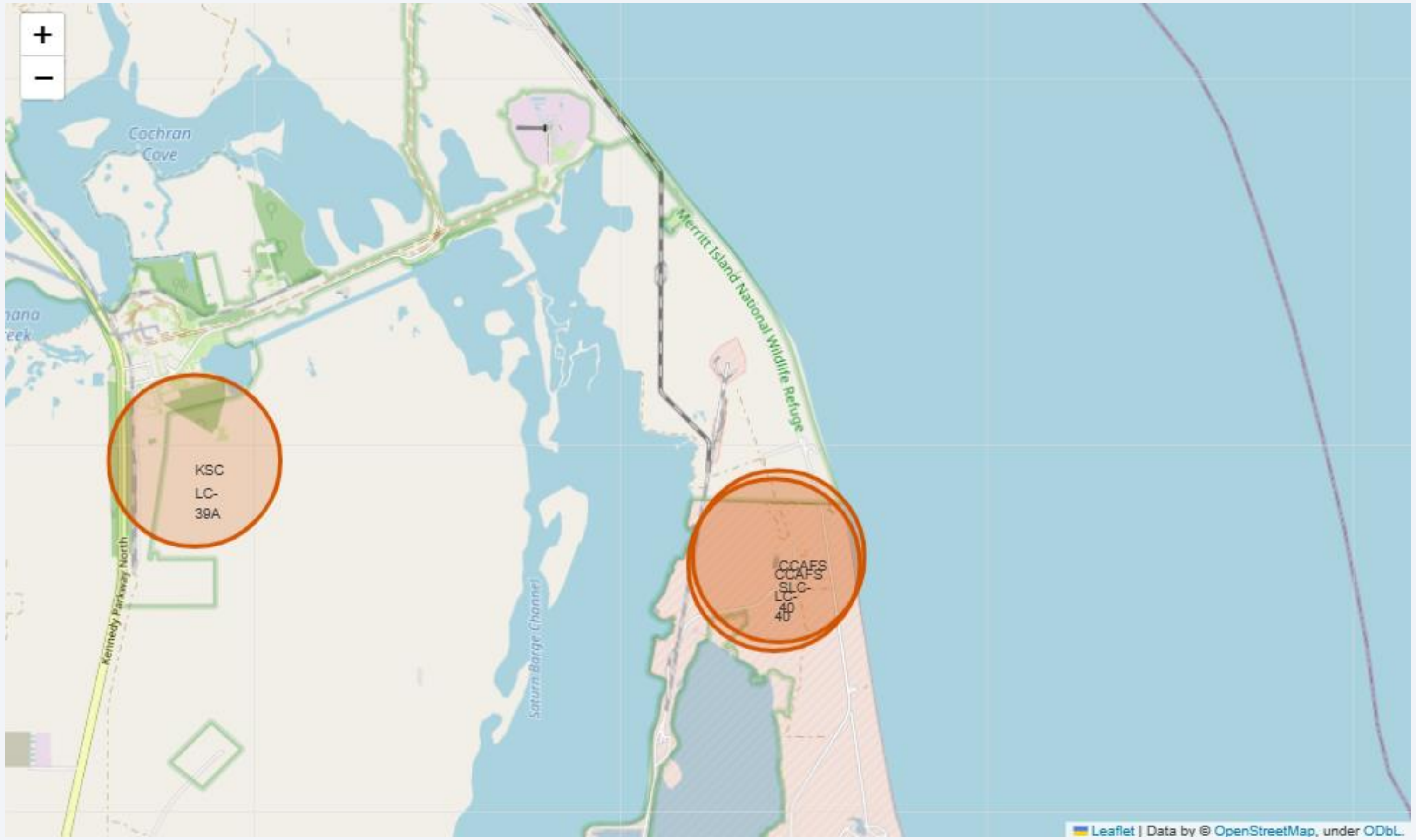
Outcomes:

- When we zoom in on a launch site, we can click on the launch site which will display marker clusters of successful landings or failed landings.
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



Launch Site Proximity

- The generated map shows that the selected launch site is close to a highway for the transportation of personnel and equipment. The launch site is also close to the coastlines for launch failure testing.
- The launch sites also maintain a certain distance from the cities (can be viewed in Jupyter Notebook).
- CCAFS SLC-40
    - 86 km from the nearest coastline
    - 21.96 km from the nearest railway
    - 23.23 km from the nearest city
    - 26.88 km from the nearest highway

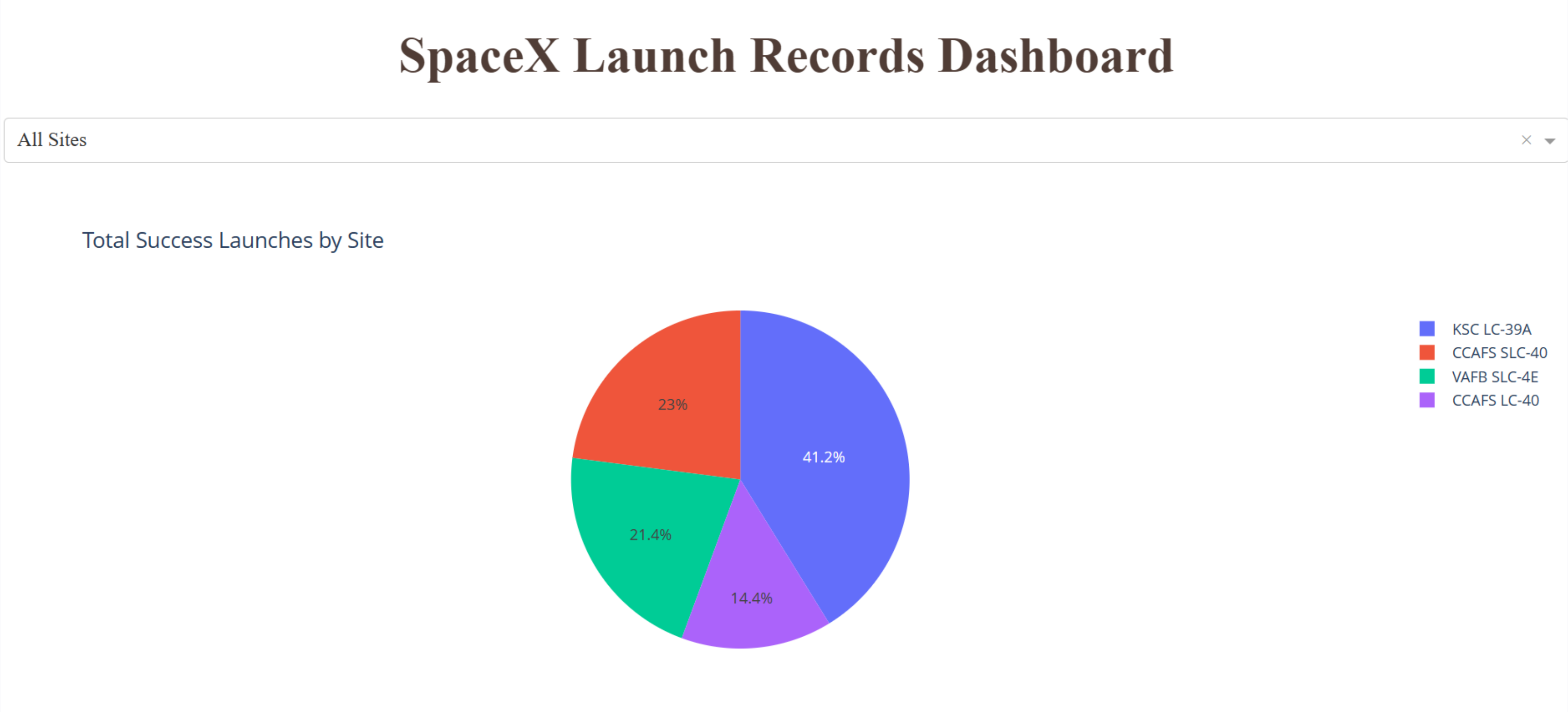


Section 4

Build a Dashboard with Plotly Dash

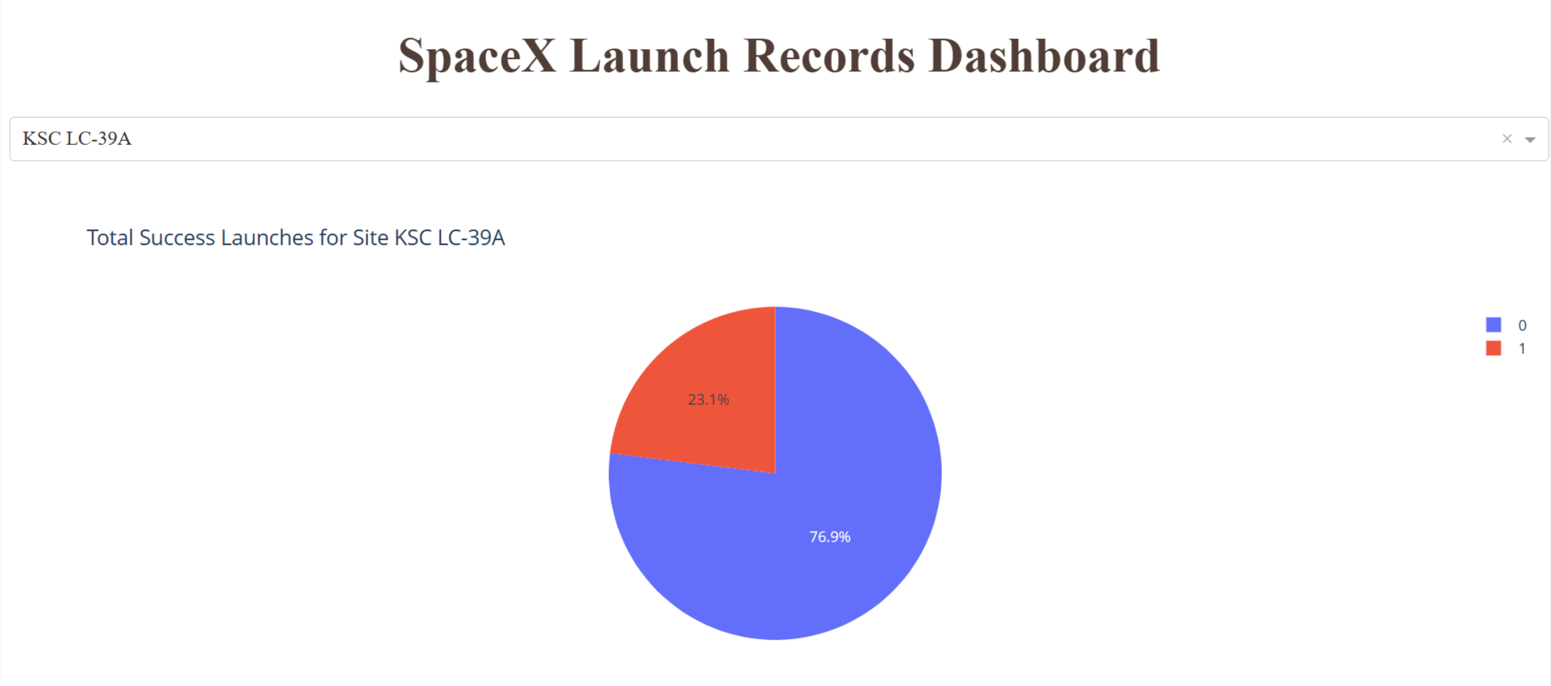
Total Successful Site Launch

- The KSC LC-39A Launch site has the most successful launches which constitutes 41% of all successful site launches.



Highest Success Ratio Launch Site

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches



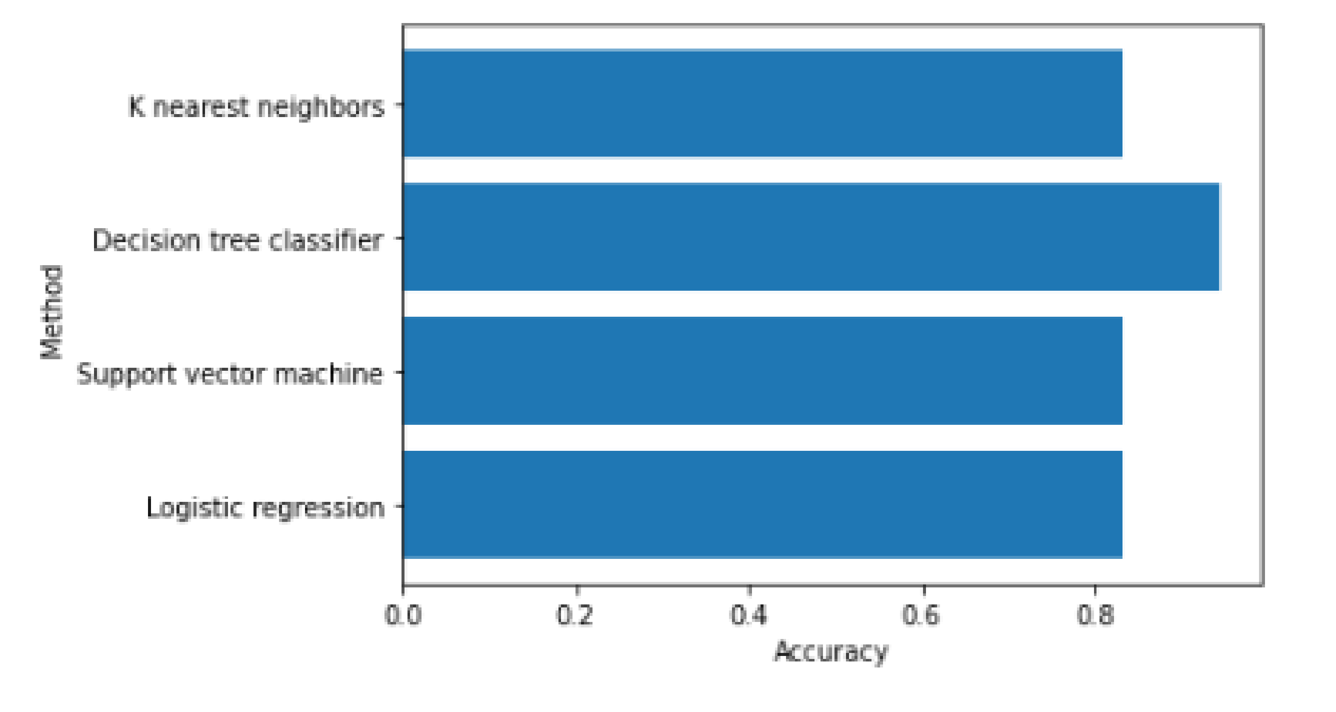
Payload Mass and Success

- 1 indicating a successful outcome and 0 indicating an unsuccessful 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.
- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- The booster version that has the largest success rate, in both weight ranges is the v1.1.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

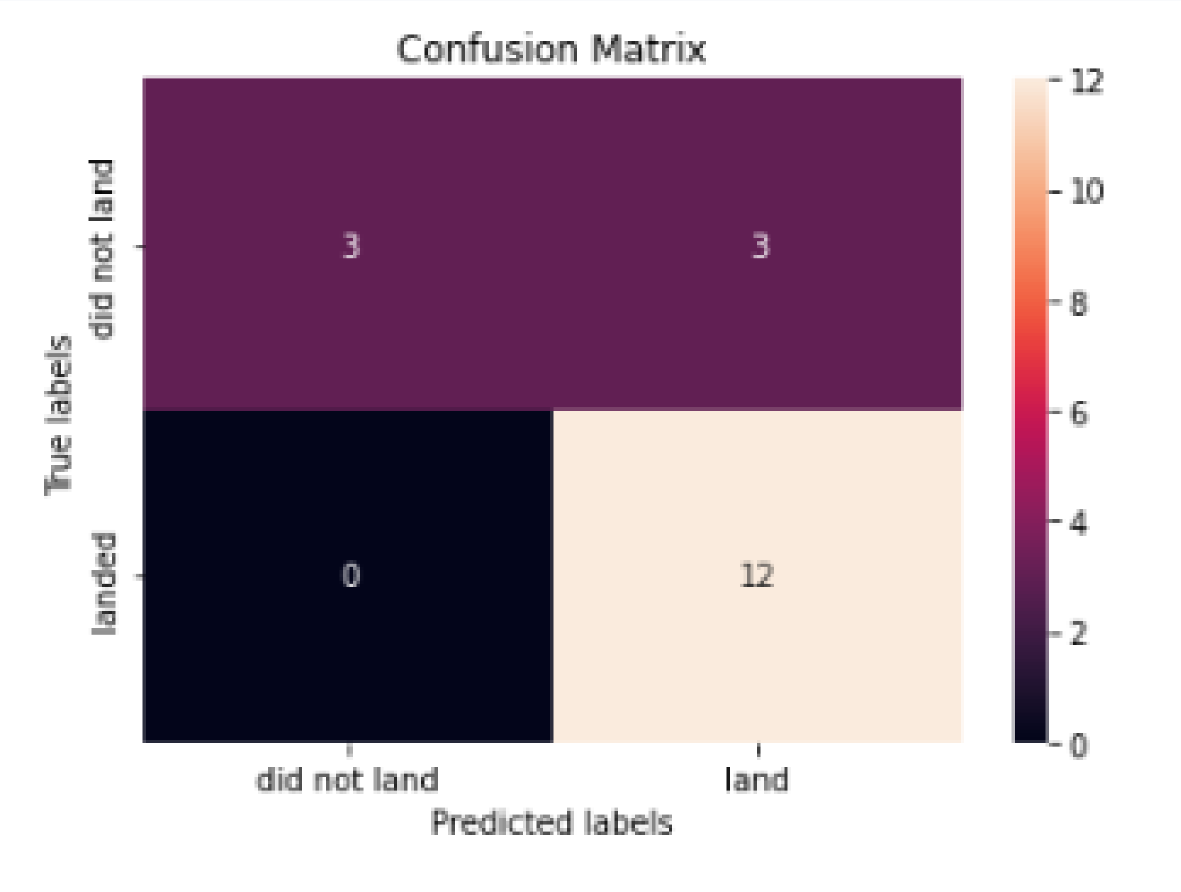


	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333

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

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 a 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

- **Model Performance:** The predictive models displayed similar performance on the test set, with the decision tree model slightly outperforming the others.
- **Strategic Equator Location:** Most of SpaceX's launch sites are strategically positioned near the equator, leveraging Earth's rotational speed to reduce fuel and booster costs.
- **Proximity to Coastlines:** All SpaceX launch sites are located in close proximity to coastlines, facilitating efficient logistics and support for launch activities.
- **Improving Launch Success:** Analysis of historical data revealed a positive trend in launch success rates, reflecting SpaceX's continuous improvement efforts.
- **KSC LC-39A Excellence:** KSC LC-39A emerged as the top-performing launch site, achieving a remarkable 100% success rate for payloads weighing less than 5,500 kg.
- **Reliable Orbits:** ES-L1, GEO, HEO, and SSO orbits demonstrated a perfect 100% success rate, making them ideal choices for high-reliability missions.
- **Payload Mass Impact:** Across all launch sites, there was a clear correlation between payload mass and launch success, highlighting the importance of payload selection and mission planning.

In summary, our research provides valuable insights for SpaceX's mission planning, site selection, and continued success in space exploration and commercial spaceflight.

Future Considerations:

- **Dataset Expansion:** Utilizing a larger dataset could enhance the reliability and generalizability of predictive analytics results.
- **Feature Analysis and PCA:** Exploring additional feature analysis or applying principal component analysis (PCA) might lead to improvements in model accuracy and reveal hidden patterns.
- **XGBoost Implementation:** The powerful XGBoost model, which was not considered in this study, could be a potential candidate for outperforming other classification models. Future analysis should include XGBoost as a model option to assess its predictive capabilities.

These considerations provide a roadmap for refining and extending the analysis, ultimately contributing to more robust predictive models and insights for SpaceX's mission planning and launch success..

Appendix

- GitHub Repository: https://github.com/evansbuycosansolis/IBM_DataScience_Capstone/

Thank you!