IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hema Bhupathi
09/07/2024

Outline

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

Executive Summary

Summary of methodologiesData collection

- Data wrangling
- EDA with Analysis
- EDA with SQL
- Building a Dashboard with Plotly Dash
- Predective Analysis

Summary of all resultsExploratory Data Analysis Result

- Interactive Analytics
- Predictive Analysis

Introduction

- Project background and contextSpaceX stands as a preeminent figure in the commercial space sector, revolutionizing space travel through affordability. Advertised on their website, Falcon 9 rocket launches boast a cost of $62 million, a significantly lower figure compared to other providers' prices, which can soar upward of $165 million per launch. This notable cost difference is primarily attributed to SpaceX's groundbreaking ability to reuse the first stage of their rockets. Thus, discerning the likelihood of a successful first stage landing becomes pivotal in calculating the overall launch cost. Leveraging publicly available data and employing machine learning models, our objective is to forecast the probability of SpaceX reusing the first stage.

- Problems you want to find answersOur endeavor aims to dissect various facets surrounding the successful or failed landing of Falcon 9's first stage. Key questions guiding our investigation include unraveling the defining characteristics of both successful and unsuccessful landings. Additionally, we seek to elucidate the intricate interplay between rocket variables such as payload mass, launch site, number of flights, and orbits in influencing the outcome of a landing. Ultimately, we endeavor to uncover the optimal conditions that would bolster SpaceX's landing success rate, thereby shedding light on crucial insights for potential competitors vying for a stake in the rocket launch market.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

    Collection of data is done through two main avenues: utilizing the SpaceX REST API and employing web scraping techniques from Wikipedia.

- Perform data wrangling

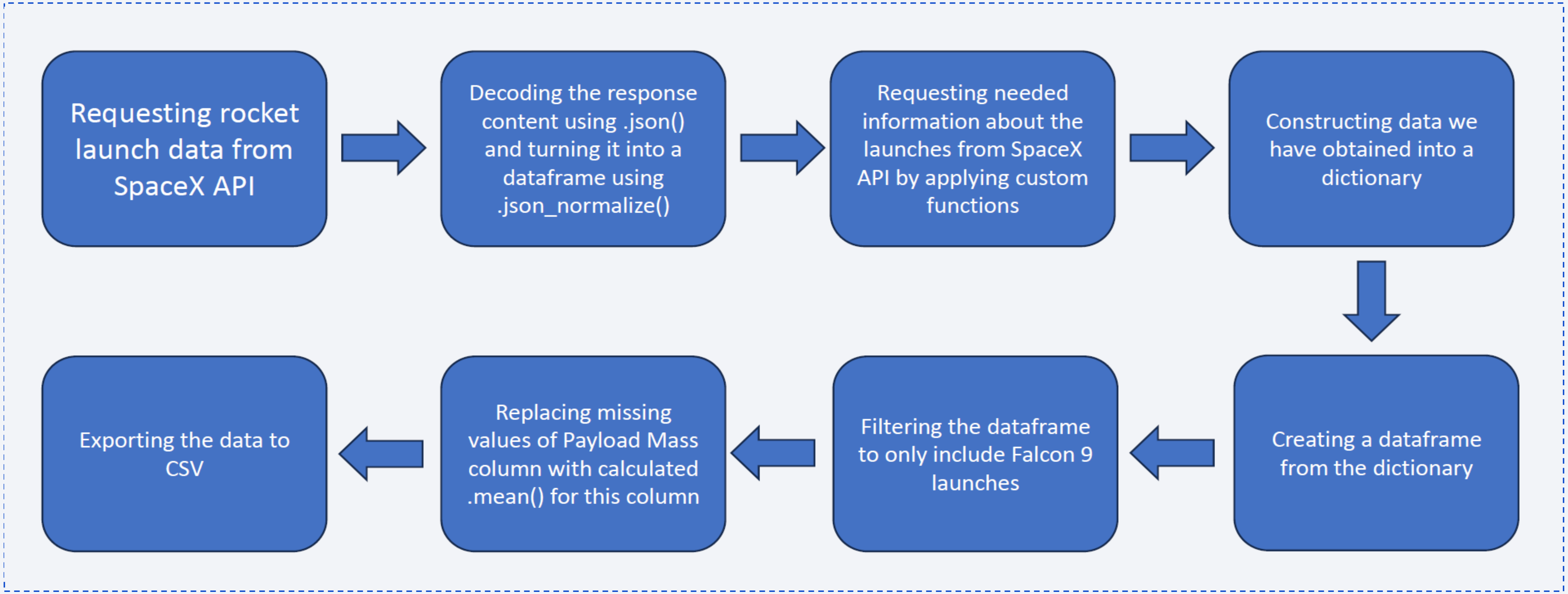
    Utilization of One Hot Encoding technique aids in transforming categorical variables into a suitable format for binary classification.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - How to build, tune, evaluate classification models

Data Collection

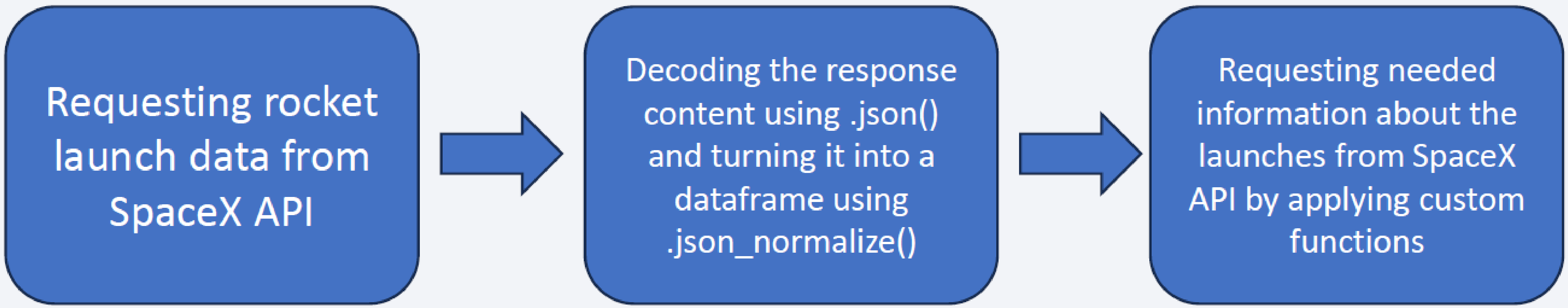
- The data collection process involved multiple methods:
- Initially, data collection was conducted using the SpaceX API.
- Subsequently, the response content was decoded into JSON format using the .json() function call. This JSON data was then normalized into a structured panda data frame using .json_normalize().
- Following data acquisition, a thorough data cleaning process was initiated. This involved identifying and addressing missing values within the dataset. Missing values were filled in where necessary to ensure data completeness and integrity.
- Additionally, web scraping techniques were employed to gather Falcon 9 launch records from Wikipedia. Beautiful Soup, a Python library, was utilized for this purpose, enabling the extraction of relevant data from the web page.

Data Collection – SpaceX API



- https://github.com/HemaBhupathi/DataScienceEcosystem/blob/Assign/jupyter-labs-spacex-data-collection-api.ipynb

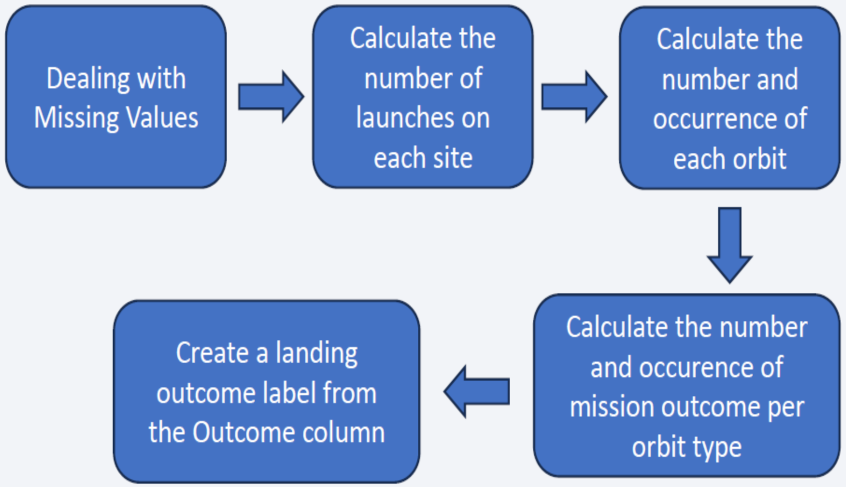
Data Collection - Scraping



- https://github.com/HemaBhupathi/DataScienceEcosystem/blob/Assign/jupyter-labs-webscraping.ipynb

Data Wrangling

- The dataset includes various scenarios where the booster did not land successfully. These scenarios encompass different outcomes based on the landing attempts. For instance, "True Ocean" indicates a successful landing in a specific region of the ocean, while "False Ocean" signifies an unsuccessful landing in the ocean. Similarly, "True RTLS" denotes a successful ground pad landing, whereas "False RTLS" indicates an unsuccessful ground pad landing. Additionally, "True ASDS" represents a successful landing on a drone ship, while "False ASDS" indicates an unsuccessful landing on a drone ship.
- To simplify and standardize these outcomes for analysis, we convert them into training labels as follows: "1" denotes a successful booster landing, while "0" signifies an unsuccessful landing. This conversion allows for a uniform representation of the landing outcomes, facilitating further analysis and modeling.
- https://github.com/HemaBhupathi/DataScienceEcosystem/blob/Assign/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- **Data Visualization Overview:**
- **Scatter Graphs:**
- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Orbit vs. Flight Number
- Payload vs. Orbit Type
- Orbit vs. Payload Mass

- **EDA with Data Visualization:**
- - Bar Graph: Success rate vs. Orbit
- - Line Graph: Success rate vs. Year

https://github.com/HemaBhupathi/DataScienceEcosystem/blob/Assign/edadataviz.ipynb

- Key Points:
- Scatter plots show correlations between variables.
- Bar graphs highlight relationships between numeric and categorical variables.
- Line graphs illustrate data trends over time for predictive analysis.

EDA with SQL

- Displaying the names of unique launch sites.
- Displaying 5 records where launch sites start with 'CCA'.
- Total payload mass carried by NASA (CRS) boosters.
- Average payload mass carried by F9 v1.1 boosters.
- Date of the first successful landing on a ground pad.
- Names of boosters with successful drone ship landings and payload mass between 4000 and 6000.
- Total number of successful and failed mission outcomes.
- Booster versions that carried the maximum payload mass.
- Failed landing outcomes in drone ships, including booster versions and launch site names, for 2015.
- Ranking of successful landing outcomes between June 4, 2010, and March 20, 2017, in descending order.

- https://github.com/HemaBhupathi/DataScienceEcosystem/blob/Assign/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map Center: The map is centered on NASA Johnson Space Center in Houston, Texas, serving as the starting location.
- Markers of Launch Sites: Red circles with labels are added at the coordinates of all launch sites, showcasing their geographical locations. The markers also indicate the proximity of the launch sites to the Equator and coastlines.
- Markers of Launch Outcomes: Marker clusters are employed to group points, allowing for the display of successful (green) and failed (red) launches at each launch site. This provides insight into which launch sites have relatively high success rates.
- Distances to Proximities: Colored lines are added to illustrate the distances between launch sites and key locations such as railways, highways, coastlines, and closest cities. This visual representation enhances understanding of the spatial relationships between launch sites and their surroundings.
- Additional Features: Popup labels, text labels, and div icons are utilized to provide additional information and enhance the user experience. These features include labels displaying the names of launch sites, successful and unsuccessful landing outcomes, and distances to key locations.
- Purpose: The interactive map serves as a comprehensive tool for visualizing launch site locations, launch outcomes, and distances to proximities. It offers valuable insights into the

https://github.com/HemaBhupathi/DataScienceEcosystem/blob/Assign/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Dashboard Components:

- Launch Sites Dropdown List: Users can select launch sites from a dropdown list, enabling focused analysis on specific sites or viewing data collectively for all sites.
- Pie Chart: A pie chart visually represents launch success, showing the total successful launches for all sites. When a specific launch site is chosen from the dropdown list, the pie chart dynamically updates to display success versus failure counts for that site.
- Slider of Payload Mass Range: Users can adjust a slider to select a payload mass range, allowing for customized analysis based on payload specifications.
- Scatter Chart: The scatter chart illustrates the correlation between payload mass and launch success across different booster versions, providing insights into how varying payload masses affect success rates.

Other Features:

- Dropdown Component: Users can choose a launch site or view data for all launch sites using the dropdown menu.
- Pie Chart: Provides a clear visualization of total success and failure counts for selected launch site(s).
- Range Slider: Allows users to select payload mass within a specified range for targeted analysis.
- Scatter Chart: Shows the relationship between payload mass and success rate, facilitating data exploration and understanding.
- https://github.com/HemaBhupathi/DataScienceEcosystem/blob/Assign/spacex_dash_app.ipynb

Predictive Analysis (Classification)

- Data Preparation:
- Load the dataset from provided URLs
- Normalize the data using StandardScaler
- Split the data into training and test sets

- Model Preparation:
- Select machine learning algorithms: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN)
- Set hyperparameters for each algorithm using GridSearchCV to find the best combination

- Model Training:
- Train each GridSearchCV model with the training dataset

- Model Evaluation:
- Get the best hyperparameters for each type of model
- Compute the accuracy of each model using the test dataset
- Plot the confusion matrix for each model

- Model Comparison:
- Compare the accuracy of all models
- Choose the model with the best accuracy
- https://github.com/HemaBhupathi/DataScienceEcosystem/blob/Assign/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

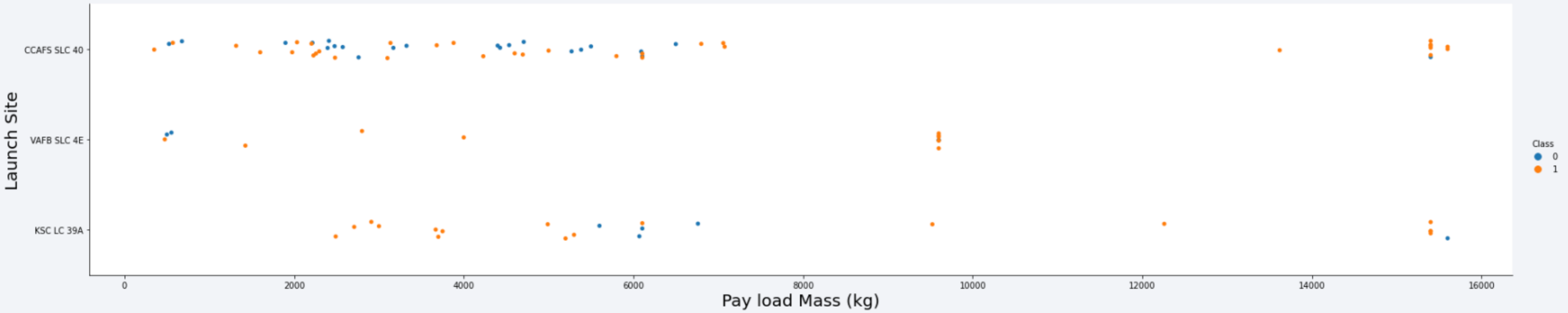
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

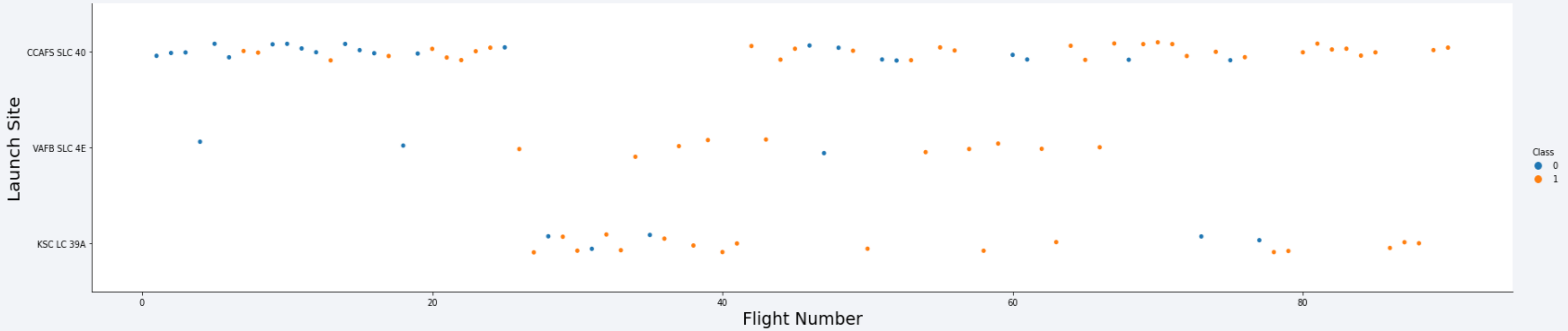
Insights drawn from EDA

Flight Number vs. Launch Site



- **Explanation:**
- CCAFS SLC 40 is responsible for nearly half of all launches.
- Higher success rates are observed at VAFB SLC 4E and KSC LC 39A.
- The initial flights experienced failures, whereas the most recent ones achieved success.
- It's plausible to assume that success rates improve with each subsequent launch.

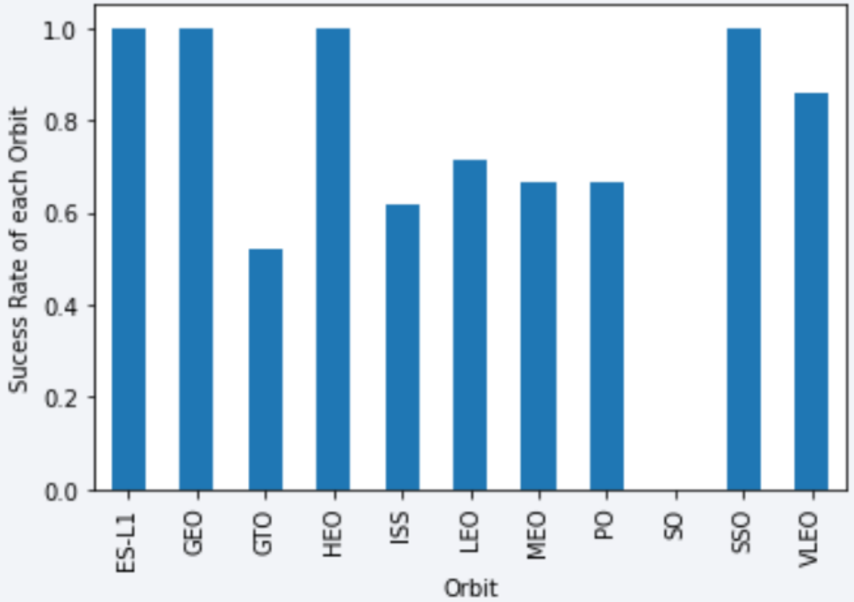
Payload vs. Launch Site



Explanation:

- Payload Mass and Success Rate: There is a direct relationship observed between payload mass and launch success rate across all launch sites. Generally, higher payload masses tend to correspond with higher success rates, indicating that launch sites are well-equipped to handle heavier payloads, resulting in increased mission success probabilities.
- Success of Launches with Payload Mass over 7000 kg: A significant proportion of launches with payload masses exceeding 7000 kg were successful, indicating the launch sites' capability to handle and successfully execute missions with heavier payloads.
- High Success Rate at KSC LC 39A: KSC LC 39A boasts an impressive 100% success rate for launches with payload masses under 5500 kg. This highlights the reliability and effectiveness of the launch site's infrastructure and operations, particularly for missions with smaller to moderate payload masses.

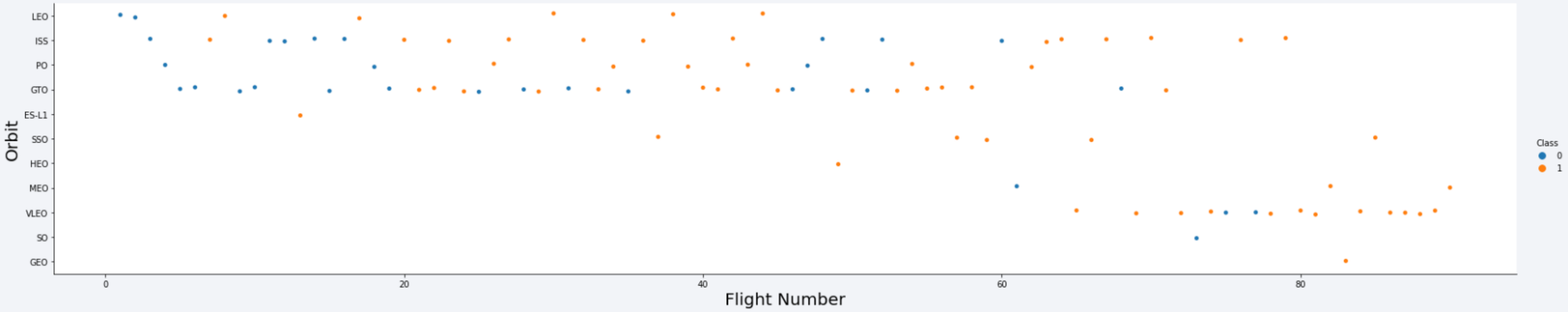
Success Rate vs. Orbit Type



Explanation:

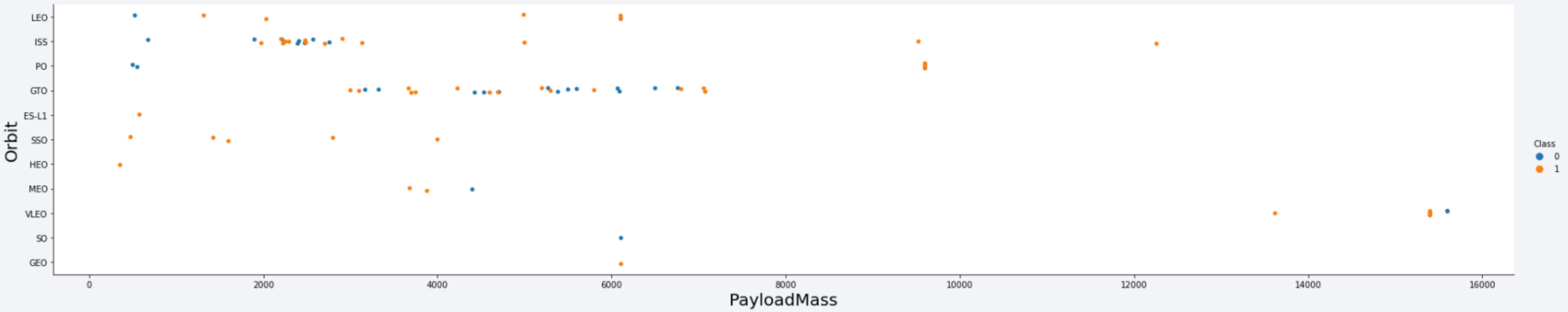
- Orbits with a perfect success rate include ES-L1, GEO, HEO, and SSO.
- Orbits with no successful missions are limited to SO.
- Orbits with success rates ranging from 50% to 85% encompass GTO, ISS, LEO, MEO, and PO.

Flight Number vs. Orbit Type



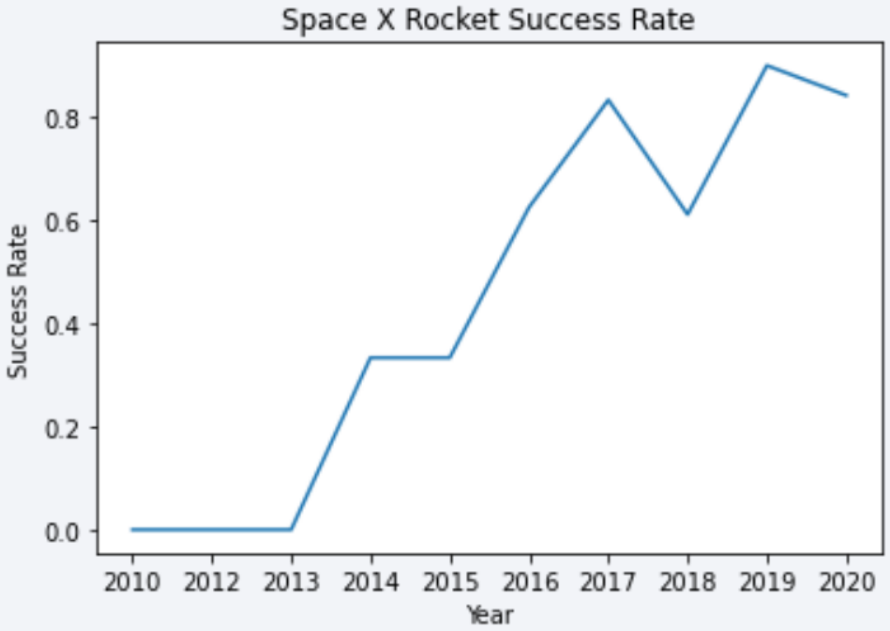
- We observe a positive correlation between the success rate and the number of flights for the LEO orbit. Conversely, for orbits like GTO, there appears to be no discernible relationship between the success rate and the number of flights. However, it's reasonable to infer that the high success rates observed for orbits such as SSO or HEO may be attributed to knowledge gained from previous launches for other orbits.

Payload vs. Orbit Type



- The weight of payloads can significantly impact the success rate of launches in specific orbits. For instance, in the LEO orbit, heavier payloads tend to enhance the success rate. Conversely, reducing the payload weight for a GTO orbit has been found to improve launch success.

Launch Success Yearly Trend



- The success rate has shown a consistent increase from 2013 to 2020.

All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[4]:
```

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

- Presenting the names of the distinct launch sites involved in space missions..

Launch Site Names Begin with 'CCA'

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

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

- Displaying five records where launch sites initiate with the string 'CCA'.

Total Payload Mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[6]:
```

total_payload_mass
45596

- Showing the cumulative payload mass transported by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[7]:
```

average_payload_mass
2534

- Indicating the mean payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[8]:
```

first_successful_landing
2015-12-22

- Listing the date of the inaugural successful landing outcome on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[9]:
```

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

- Enumerating the booster names that achieved success in drone ship landings with payload masses greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

- Enumerating the overall count of successful and failed mission outcomes.

Boosters Carried Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[11]:
```

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

- Displaying the names of booster versions that transported the maximum payload mass.

2015 Launch Records

In [12]:
```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Listing instances of failed landing outcomes on drone ships, alongside their booster versions and launch site names for the months in the year 2015.

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

In [13]:
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

landing__outcome	count_outcomes
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

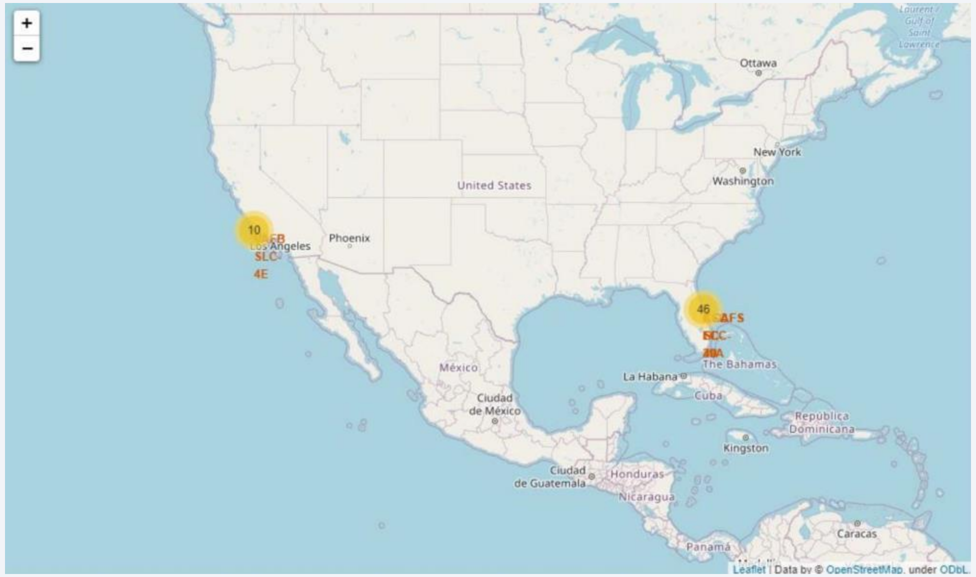
- Ranking the frequency of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between June 4, 2010, and March 20, 2017, in descending order.

Section 3

Launch Sites Proximities Analysis

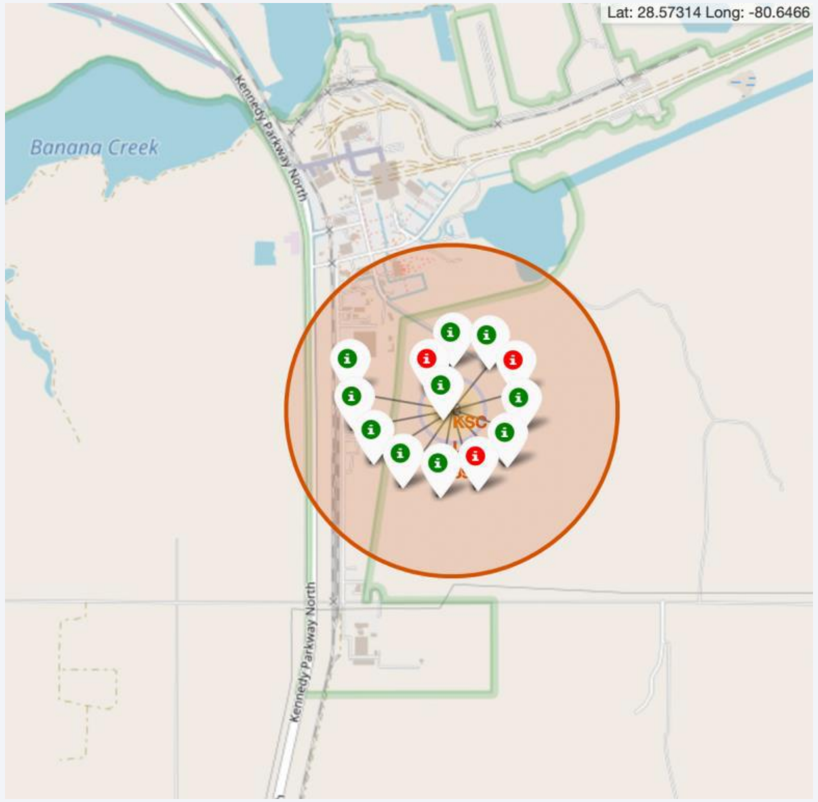
Ground Station

- Most launch sites are strategically situated near the Equator line. This positioning takes advantage of the Earth's rapid rotation at the equator, with speeds reaching 1670 km/hour. Launching from the equator allows spacecraft to leverage this inertia, aiding in maintaining orbital velocities.
- Additionally, all launch sites are located in close proximity to coastlines. Launching rockets over the ocean minimizes the risk of debris falling or exploding near populated areas, ensuring safety during launches.



Color launch records

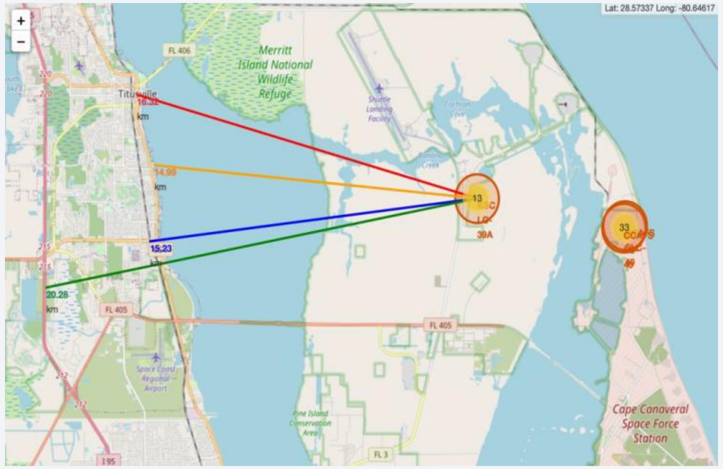
- Color-labeled markers facilitate the easy identification of launch sites with relatively high success rates.
- Green Marker = Successful Launch
- Red Marker = Failed Launch
- Notably, a certain launch site exhibits a notably high success rate, as indicated by the markers.



Distance from the launch site to proximity

- **Explanation:**
- Visual analysis reveals that the launch site is situated:
- relatively close to a railway (15.23 km)
- relatively close to a highway (20.28 km)
- relatively close to a coastline (14.99 km)

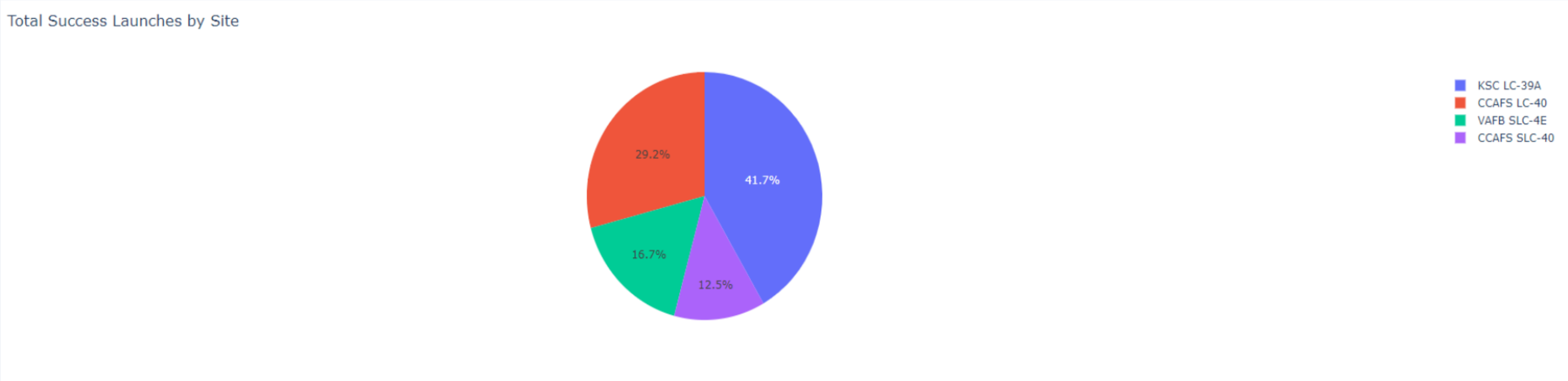
- Additionally, the launch site is in close proximity to its nearest city (16.32 km).
- Considering the high speeds of failed rockets, covering distances of 15-20 km within seconds, the proximity to populated areas poses potential safety concerns.



Section 4

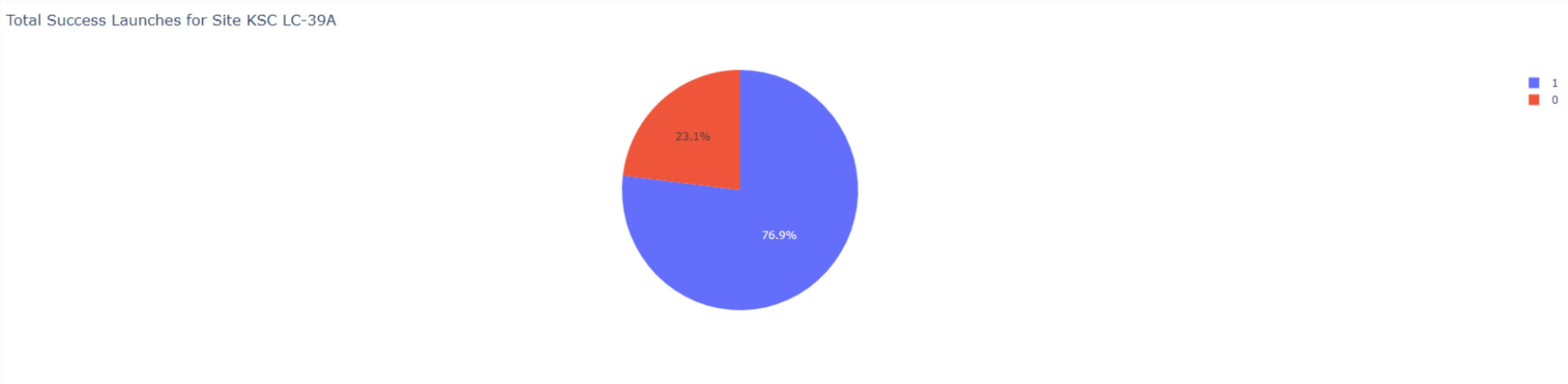
Build a Dashboard with Plotly Dash

Launch Success



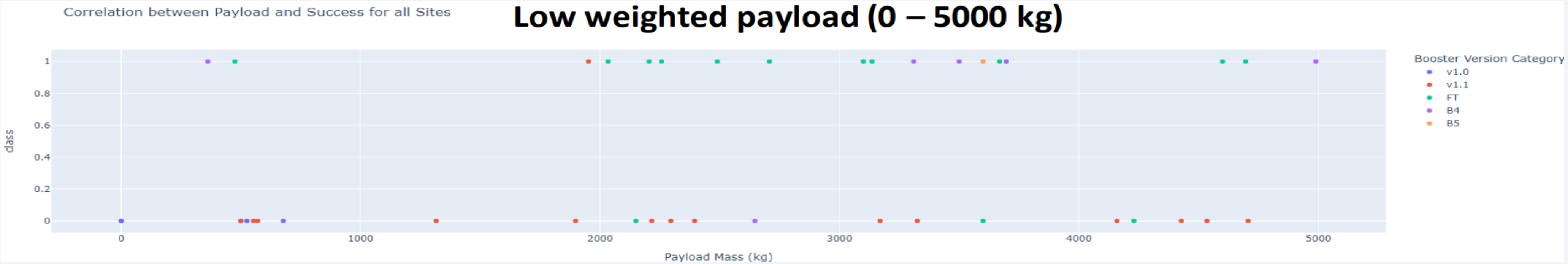
The graph displays that KSC LC-39A leads in successful launches among all sites.

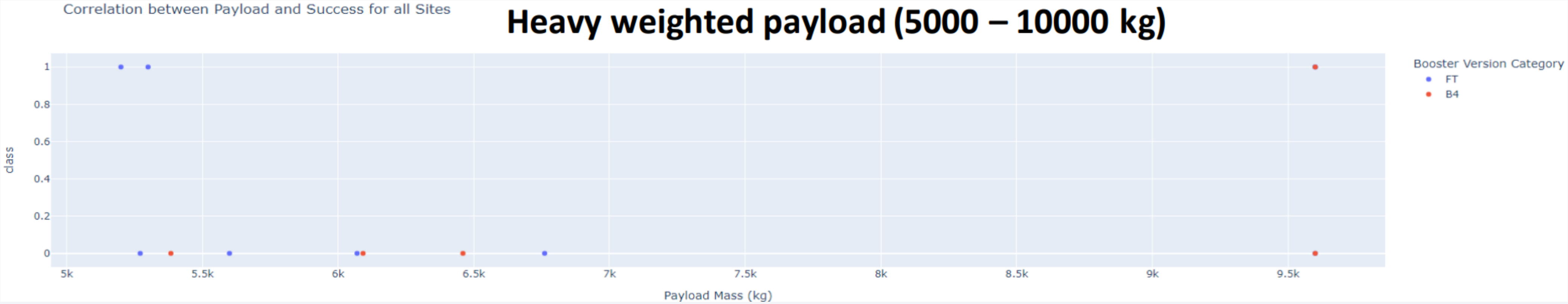
Highest Launch



- KSC LC-39A boasts the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs Launch Outcome





Low-weighted payloads have a better success rate than the heavy-weighted payloads.

Section 5

Predictive Analysis (Classification)

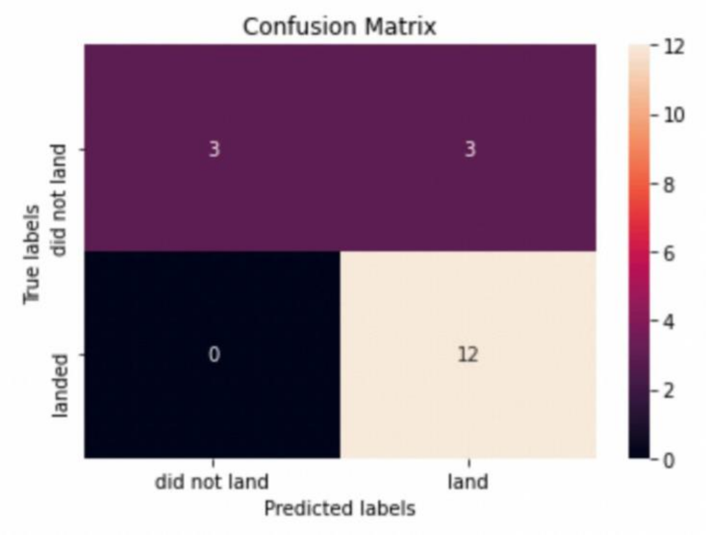
Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

- Relying solely on Test Set scores doesn't definitively establish the best-performing method.
- •Similar Test Set scores across methods may stem from the small sample size (18 samples). Consequently, all methods underwent evaluation using the entire Dataset.
- •Examination of scores from the entire Dataset identifies the Decision Tree Model as the top performer. This model not only achieves higher scores but also maintains the highest accuracy.

Confusion Matrix



- As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models are false positives.

Conclusions

- Success in space missions depends on various factors, including the launch site, orbit, and payload mass, as well as cumulative launch experience.
- Orbits such as GEO, HEO, SSO, and ES-L1 demonstrate remarkable success rates, indicating their reliability for space missions.
- Payload mass plays a crucial role in mission success, with lighter payloads generally showing better outcomes compared to heavier ones.
- Despite similar test accuracies among models, the Decision Tree Algorithm is favored due to its superior train accuracy.
- The reasons behind disparities in launch site performance, notably KSC LC-39A's superiority, remain unclear and warrant further investigation.
- The consistent upward trend in launch success rates over the years reflects advancements in space exploration.

Appendix

- Thank you

Thank you!