IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Sheik Sufyan>
<26th September, 2025>

Executive Summary

- This project aimed to predict the success of SpaceX's Falcon 9 first-stage landings. Data was collected from the SpaceX REST API and web scraping Wikipedia. After extensive data wrangling and exploratory data analysis (EDA) using SQL and data visualization, we uncovered key factors influencing launch success, such as launch site, payload mass, and orbit type. Interactive maps and dashboards were built to visualize these relationships.
- Finally, several classification models were trained to predict landing outcomes, with the **Decision Tree** model achieving the highest accuracy .

Introduction

SpaceX's ability to reuse the Falcon 9 first stage is the main factor reducing launch cost.
A competing company needs to predict the success probability of a reusable landing to set
competitive launch prices and allocate resources (e.g. drone ship deployment).

Questions to Answer from this Project:

- Which launch characteristics (Launch Site, Orbit Type, Payload Mass) predict a successful first stage landing?
- Can a machine learning model accurately classify the landing outcome (Success vs. Failure)?

Section 1

Methodology

Data Collection Methodology

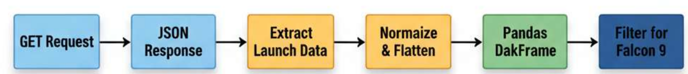
- Data was collected from the official SpaceX REST API (JSON format) and historical records scraped from Wikipedia (HTML table).
- * Perform data wrangling: Cleaned and imputed missing values (e.g., assumed `Payload Mass` missing $\\rightarrow$ 0). Crucially, we transformed the `Outcome` string descriptions into a single numerical `Class's variable (1 for success, 0 for failure) and applied One-Hot Encoding to all categorical features for ML model readiness. *Perform exploratory data analysis (EDA) using visualization and SQL: Used Matplotlib/Seaborn to plot relationships (e.g., Flight Number vs. Success Rate, Orbit vs. Success Rate). Used SQL queries to extract specific data insights, like total mass by a customer.
- * Perform interactive visual analytics using Folium and Plotly Dash: Folium was used to visualize launch site locations and landing outcomes geographically. A Plotly Dash dashboard provided an interactive tool for stakeholders to filter data by Payload Mass and Launch Site.
- * Perform predictive analysis using classification models: We split the data (train/test) and utilized Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors models.
- * `GridSearchCV` was used for systematic hyperparameter tuning on each model. Evaluation was based on the accuracy score on the test set, with the Confusion Matrix providing detailed performance metrics.

Data Collection

We sent HTTP GET requests to the SpaceX API's `/v4/launches/` endpoint. The returned JSON data was parsed to extract relevant launch details, rocket configuration, and landing outcomes, which were then normalized into a Pandas DataFrame.

Data Collection – SpaceX API

- Flowchart Key Phrases: GET Request → JSON Response → Extract Launch Data → Normalize & Flatten → Pandas DataFrame → Filter for Falcon 9.
- * GitHub URL colon https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

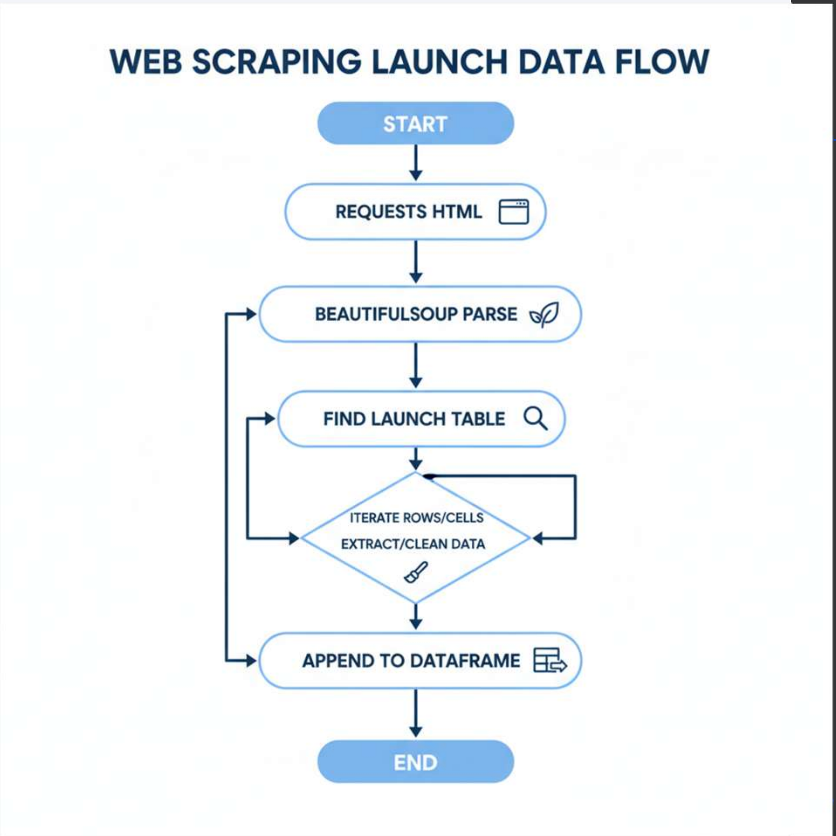


Data Collection - Scraping

Falcon 9 launch data were scraped from Wikipedia using requests and parsed with BeautifulSoup to extract Date, Launch Site, Orbit, Payload and Outcome, which were cleaned and combined with the API data in a second DataFrame.

Flowchart Key Phrases: requests HTML → BeautifulSoup Parse → Find Launch Table → Iterate Rows/Cells → Extract/Clean Data → Append to DataFrame.

GitHub URL:
https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Missing Value Imputation: Missing Payload Mass values were filled, with the mean value.
- Categorical nulls were handled by removal of rows.
- Target Variable Creation: The string Landing Outcome was converted to a simple integer label: Class (1 for success, 0 for failure).
- Feature Encoding: Categorical features like Orbit, LaunchSite, and LandingPad were transformed into a numerical format using One-Hot Encoding (creating dummy variables) to make them compatible with the ML algorithms.
- GitHub URL: https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts Plotted and Why:
- Scatter Plots (Flight Number & Payload Mass vs. Launch Site): Used to visually check for patterns of success/failure based on the rocket's mission experience (Flight Number) and size of the cargo (Payload Mass) at different geographical locations.
- Bar Chart (Success Rate by Orbit): Showed that certain orbits (e.g., ES-L1, HEO, SSO) have a perfect 100% success rate, indicating highly optimized parameters for these mission profiles.
- Line Chart (Yearly Success Rate): Confirmed the positive time trend, with success rate steadily climbing after 2015/2016, suggesting process and technology maturity.
- GitHub URL: https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/edadataviz.ipynb

EDA with SQL

- Unique Launch Sites: Found the three primary sites: CCAFS SLC 40, VAFB SLC 4E, and KSC LC 39A.
- NASA CRS Payload Mass: Calculated the total mass of cargo delivered for the NASA Commercial Resupply Services contracts.
- Average Payload Mass (F9 v1.1): Determined the typical payload capacity for a specific early booster version.
- First Successful Ground Landing: Located the date of the first recovery on land (2015-12-22), a key milestone.
- Ranking Landing Outcomes: Showed the distribution and frequency of different landing results (e.g., True ASDS, False ASDS, True RTLS) over a specific time range.
- GitHub URL: https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We created a Folium map centered on the US NASA Johnson Space Centre.
- Markers were placed at the exact coordinates of each Launch Site.
- Also added success status with markers. The markers were color-coded: Green for successful landings (Class=1) and Red for failed landings (Class=0).
- The colored markers allow for a quick visual assessment of the geographic distribution of success and failure, revealing that successful landings are concentrated at the newer sites (KSC and VAFB).
- Also created maps to show nearest Coast line, Highways and Cities etc. from Launch Site. Observed that all launch sites are near to Coast line.
- GitHub URL...: https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/Copy%20of%20WithAnswers_lab_jupyter_launch_site_location%20.ipynb

Build a Dashboard with Plotly Dash

- The dashboard includes an interactive scatter plot showing Payload Mass vs. Success, payload mass can be adjusted with slider.
- Also it includes Pie Chart showing Success rate per Launch Site., launch site can be select from drop down.
- Key interactions are a dropdown menu to filter by Launch Site and a range slider to filter by Payload Mass.
- The interactive filters allow stakeholders to dynamically test hypotheses (e.g., success rate for heavy payloads at a specific site) and confirms the relationship between Payload Mass and landing success.

Predictive Analysis (Classification)

- We Standardize the input, we split the data (train/test) and initially trained four models: Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors.
- Each model was improved and tuned using GridSearchCV to systematically find the optimal hyperparameters.
- The models were evaluated based on the accuracy score on the test set, with the Decision Tree model yielding the highest result for test set.
- Also evaluated using Confusion Mattrix, all the emodels had similar results. False positive (3) was the concerns and was same in all models.
- Flowchart Key Phrases: Standardize/Split Data → Train Initial Models → GridSearchCV (Hyperparameter Tuning) → Evaluate/Compare Accuracy /Confusion Matrix→ Select Best Model (Decision Tree)
- GitHub URL...: https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/Updated-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

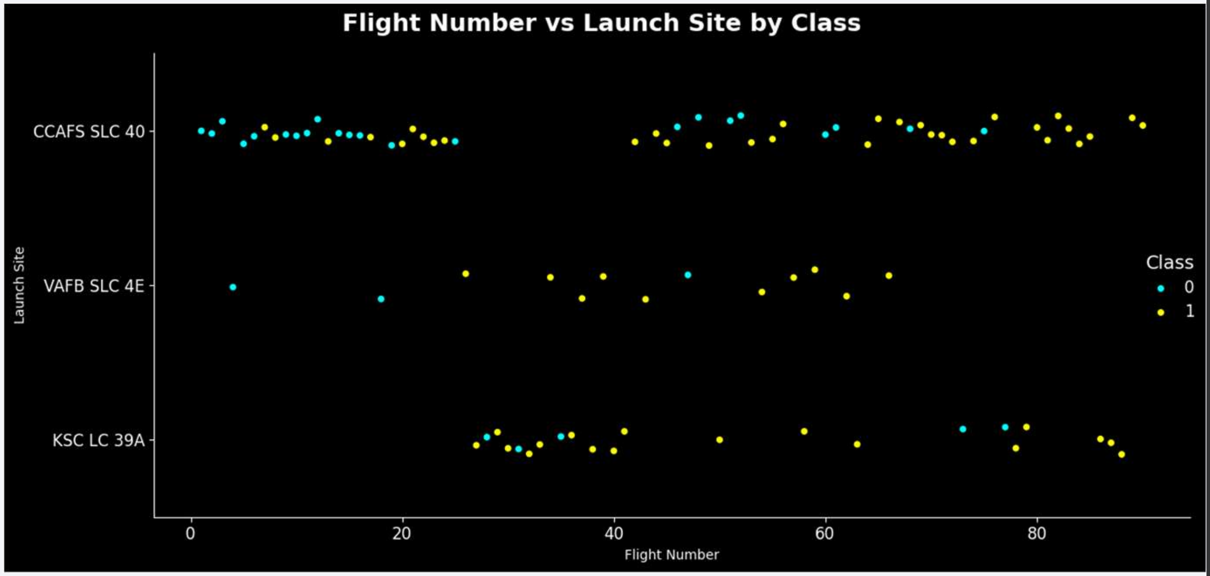
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

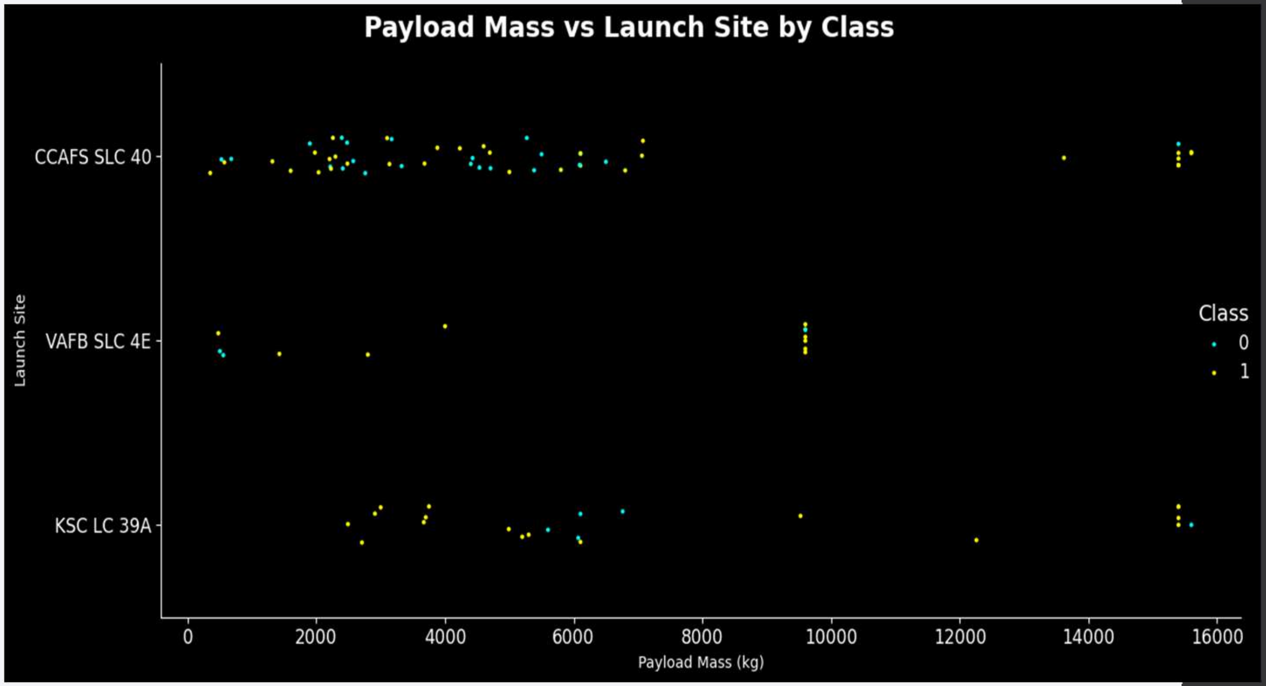
Flight Number vs. Launch Site

- \- We can observe that Success rate increases as flight number increases which indirectly tells us that flights launched later had higher success rate, indicating they corrected their errors and matured processes.
- -- Launch Site CCAFS SLC40 had high success rate.



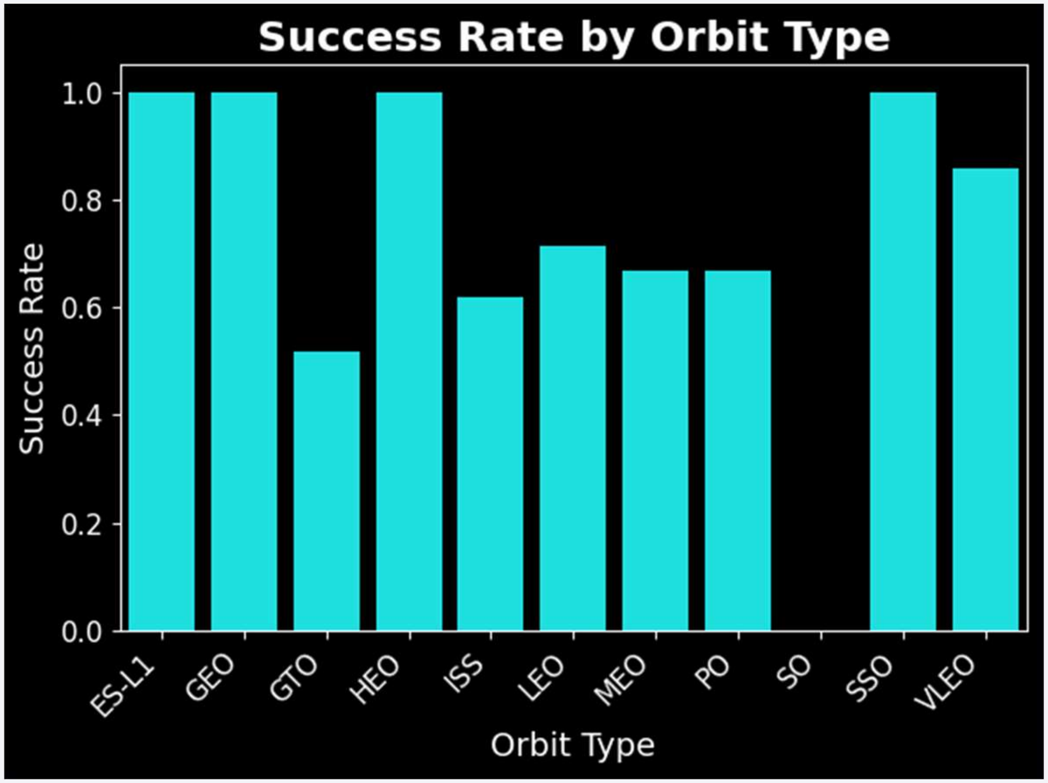
Payload vs. Launch Site

- \- For Heavy Playload mass two sites CAFS SLC40 and KSC LC 39A are used and they had good success rate.
- -Most of the launches are for pay load mass less than 10,000 KG



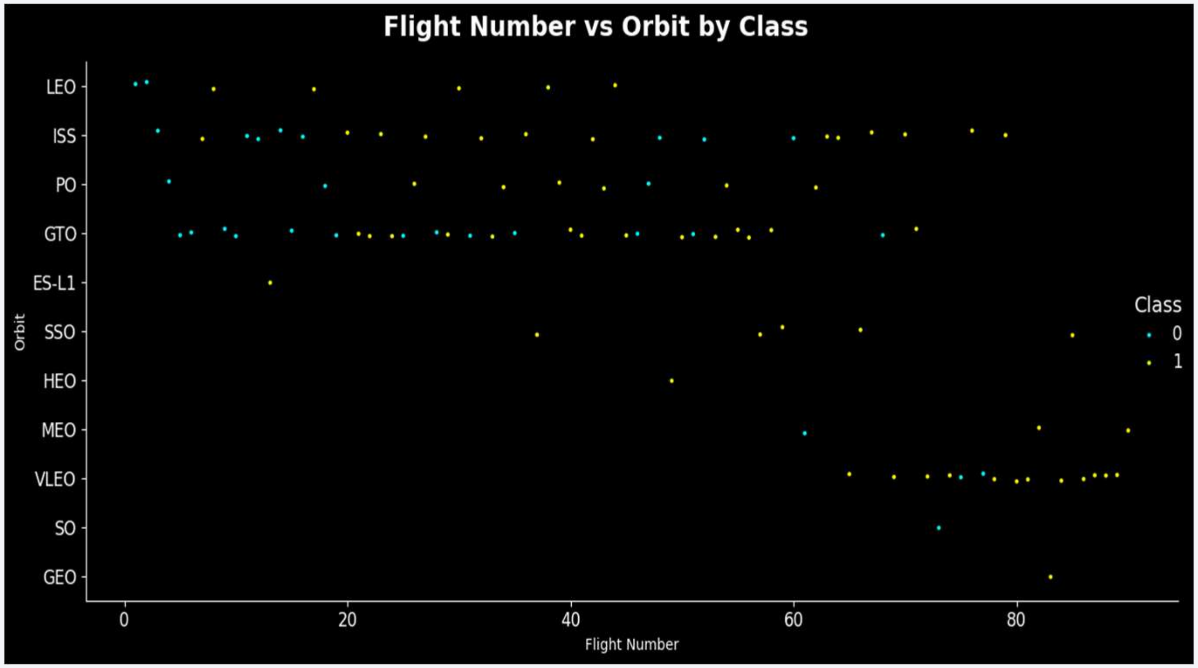
Success Rate vs. Orbit Type

- - Orbits ES L1, GTO, HEO, SSO has 100% Success rate.
- Orbit VLEO and LEO had good success rates.
- - GTO, MEO, PO Orbits had less success rate.



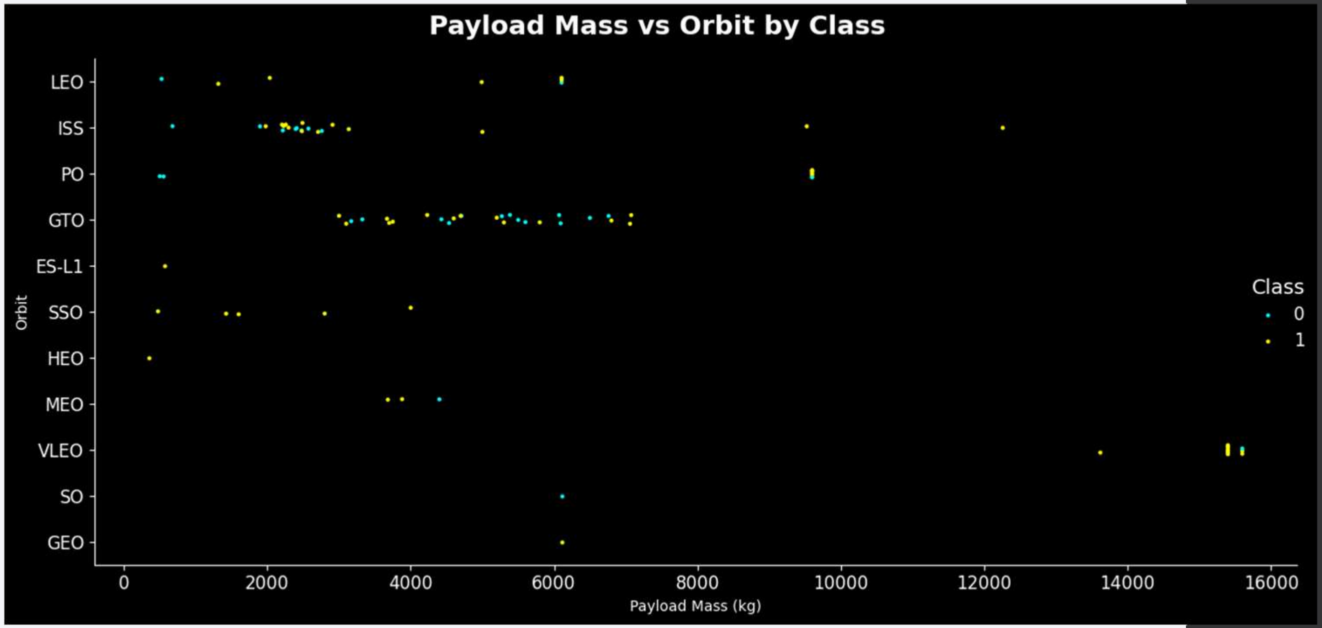
Flight Number vs. Orbit Type

- - Most of the launches are for Orbits LEO, ISS, PO
- - Success rate increased as Flight number increases.
- - Most of the initial flights were to Orbits LEO, ISS and PO but later frequency has decreased in later flights.



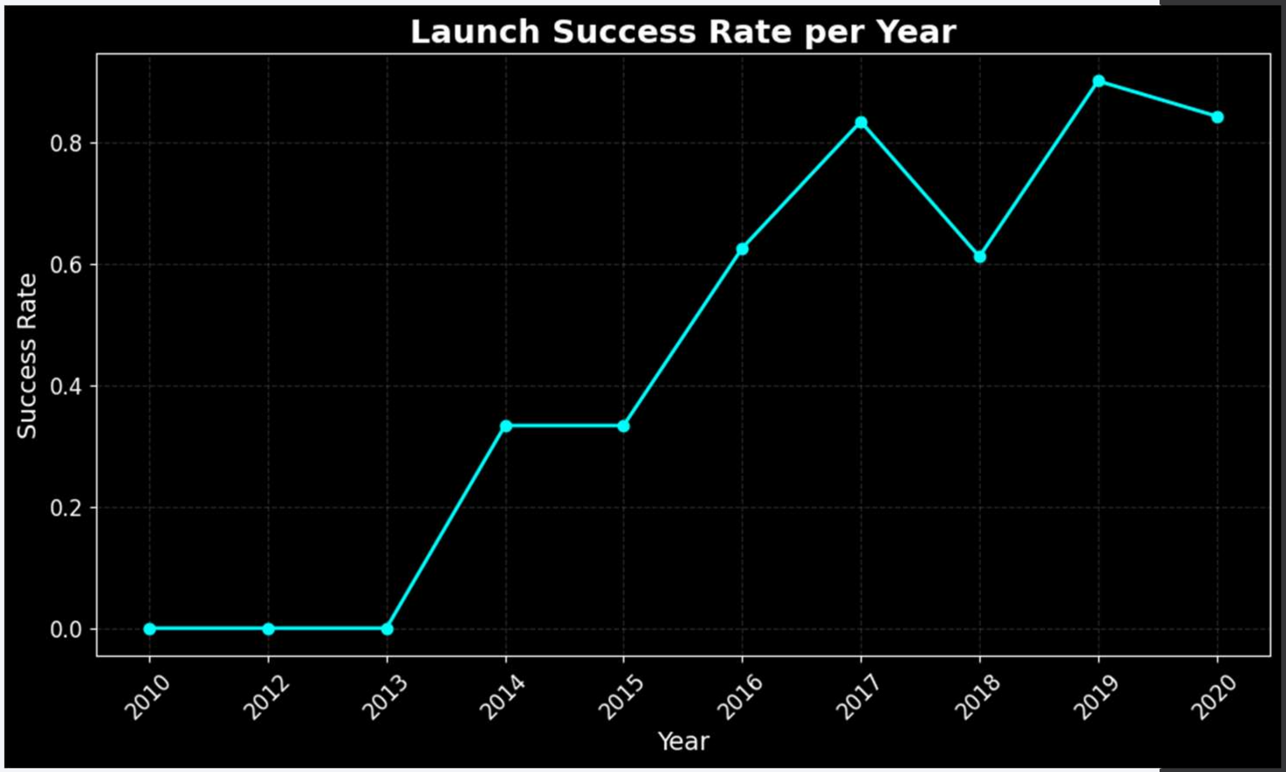
Payload vs. Orbit Type

- - Higher Payload mass flights were to Orbit VLEO and they were successful.
- - LEO had consistent success irrespective of Payload mass.



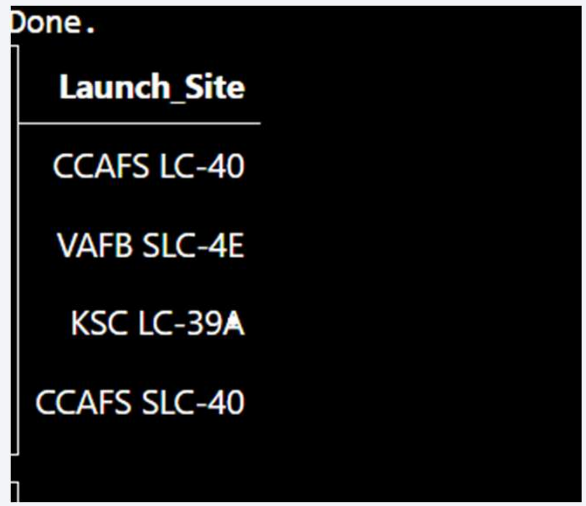
Launch Success Yearly Trend

- - Success rate increased after year 2013.
- - There is consistent increase in Success rate from 2013.
- -



All Launch Site Names

- There are four unique Launch sites used by SpaceX Falcon9.
- - CAFS LC-40
- - VAFB SLC-4E
- - KSC LC-39A
- - CAFS SLC-40



Done.

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

Launch Site Names Begin with 'CCA'

- Following are the Launch sites which begin with CCA

- - CCAFS LC-40
- - CCAFS SLC-40

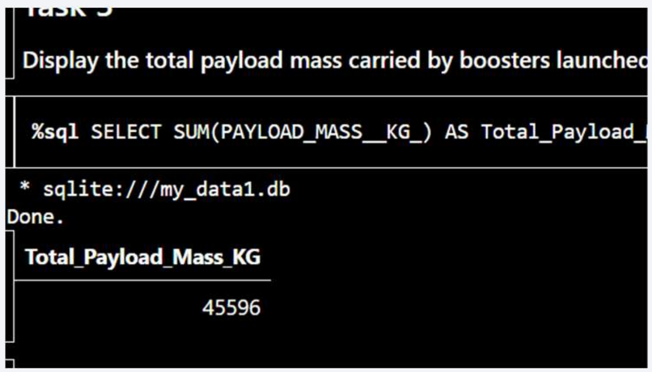
```
In [12]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

 * sqlite:///my_data1.db
Done.
Out[12]:
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success

Total Payload Mass

- The total payload carried by booste rs from NASA is 45596 KG.



Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is **2928.4 KG.**

Task 4

Display average payload mass carried by booster version F9 v1.1

In [15]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass_KG FROM SPACEXTABLE WHERE Booster_Vers
```

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

Out[15]:

Avg_Payload_Mass_KG
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 22nd December, 2015.

```
%sql SELECT MIN(Date) AS First_Successful_Ground_Pad_Landing FROM SPACEXTA
```

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

First_Successful_Ground_Pad_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are..

1. F9 FT B1022
2. F9 FT B1026
3. F9 FT B1021.2
4. F9 FT B1031.2

```
[23]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE W

 * sqlite:///my_data1.db
Done.
[23]: Booster_Version
      F9 FT B1022
      F9 FT B1026
      F9 FT B1021.2
      F9 FT B1031.2
```

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are as follows:

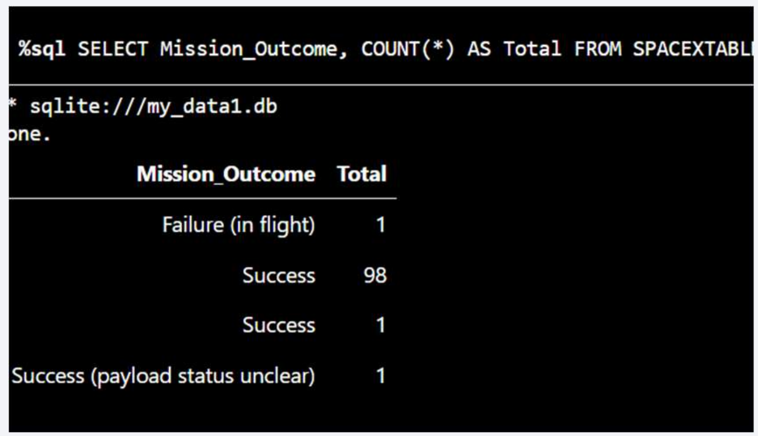
Mission_Outcome Total

Failure (in flight) 1

Success 98

Success 1

Success (payload status unclear) 1



```
%sql SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLI
* sqlite:///my_data1.db
one.
```

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are as follows:
- F9 B5 B1048.4
- F9 B5 B1049.4
- F9 B5 B1051.3
- F9 B5 B1056.4
- F9 B5 B1048.5
- F9 B5 B1051.4

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

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

Month Booster_Version Launch_Site Landing_Outcome

- January F9 v1.1 B1012 CCAFS LC-40 Failure (drone ship)

April F9 v1.1 B1015 CCAFS LC-40 Failure (drone ship)

```
%sql SELECT CASE SUBSTR(Date, 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'Feb
```

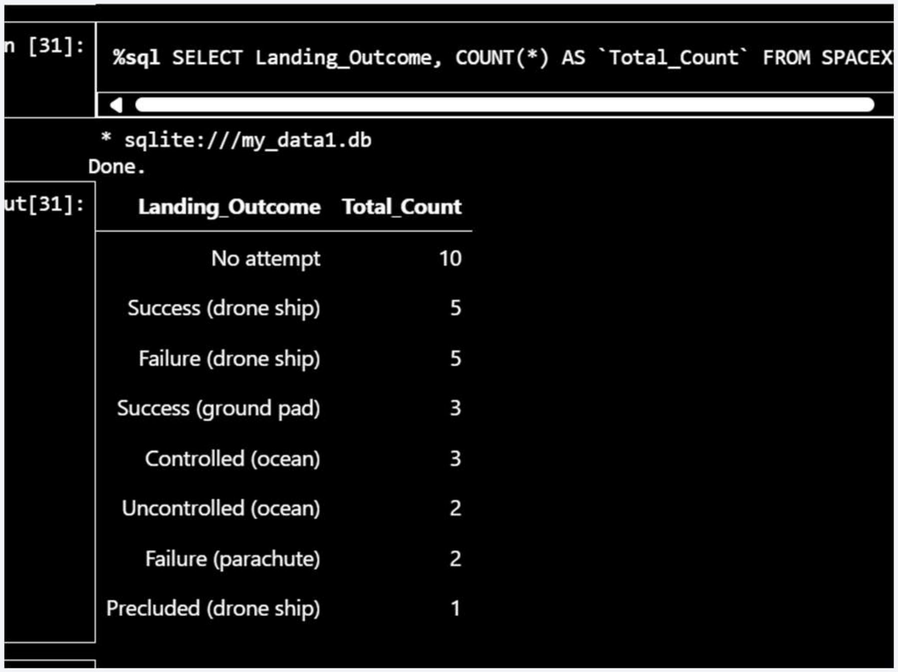
```
* sqlite:///my_data1.db
Done.
```

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

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

- Following is the 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

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



```
In [31]: %sql SELECT Landing_Outcome, COUNT(*) AS `Total_Count` FROM SPACEX
 * sqlite:///my_data1.db
Done.
```

Out[31]:

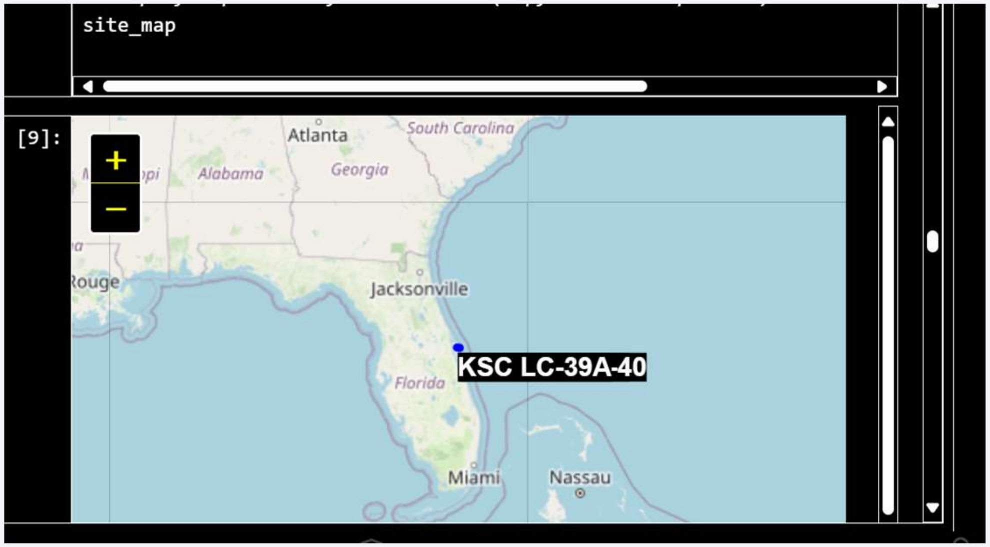
Landing_Outcome	Total_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

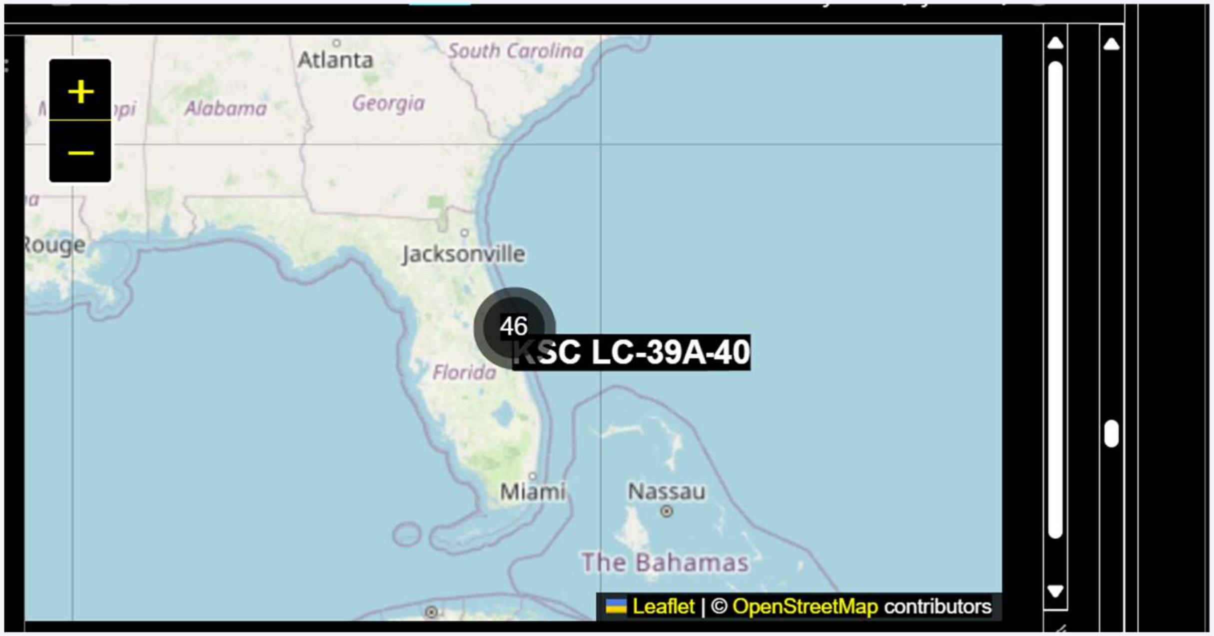
Folium Map of all Launch Locations

- Following the screenshot of all 4 Launch sites location on Map.
- They re all in Northern Hemisphere near to Equator.
- - Also they re near to Coast line.



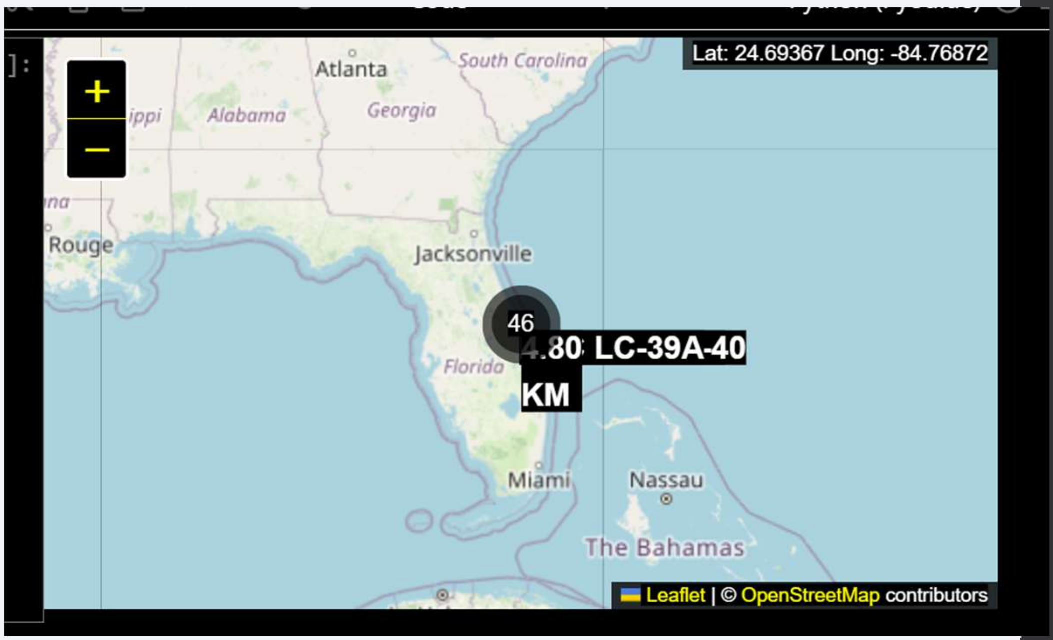
Folium Map Launch Site Success Rate

- Following Map shows Launch outcome on each launch location.
- - We can observe that KSC LC39A location had highest successful launches.



Folium Map Proximities to Launch Site

- Following map shows Launch site proximity nearest Coast line.
- - We can observer that Launch sites are near to Coastal line which facilitates fall of debris in case of accident in sea or recovery of part in sea.
- - Launch sites are fairly distant from Cities to avoid any harm to population, also there are no public railway line . There are high ways near launch site but they do not lead to launch site as they are secured areas.

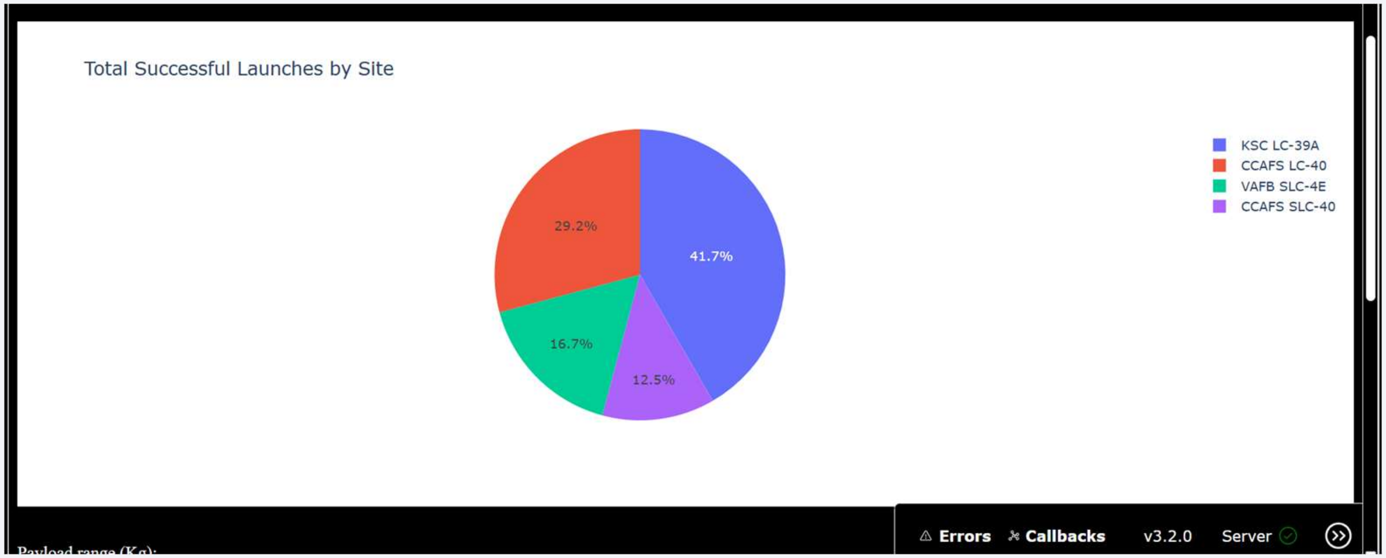


Section 4

Build a Dashboard with Plotly Dash

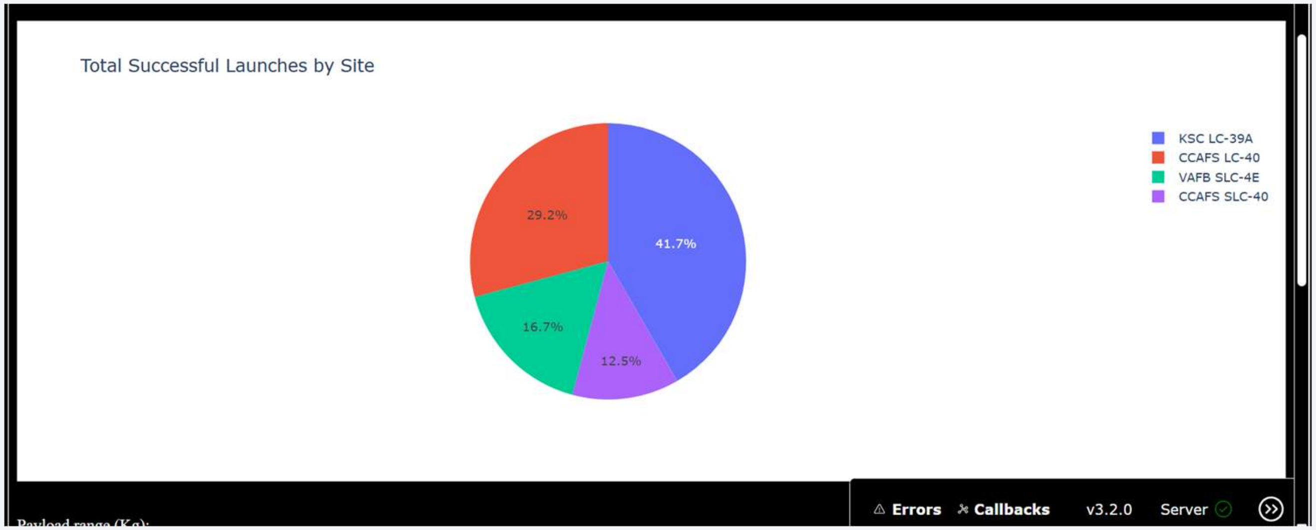
Dashboard Success Rate of Launch Sites (Pie Chart)

- Following is the Success rate of all Launch Sites, from this we can infer”
- - KSC LC-39E has Highest Success Percentage.
- - CCAFS SLC-40 had lowest success rate.



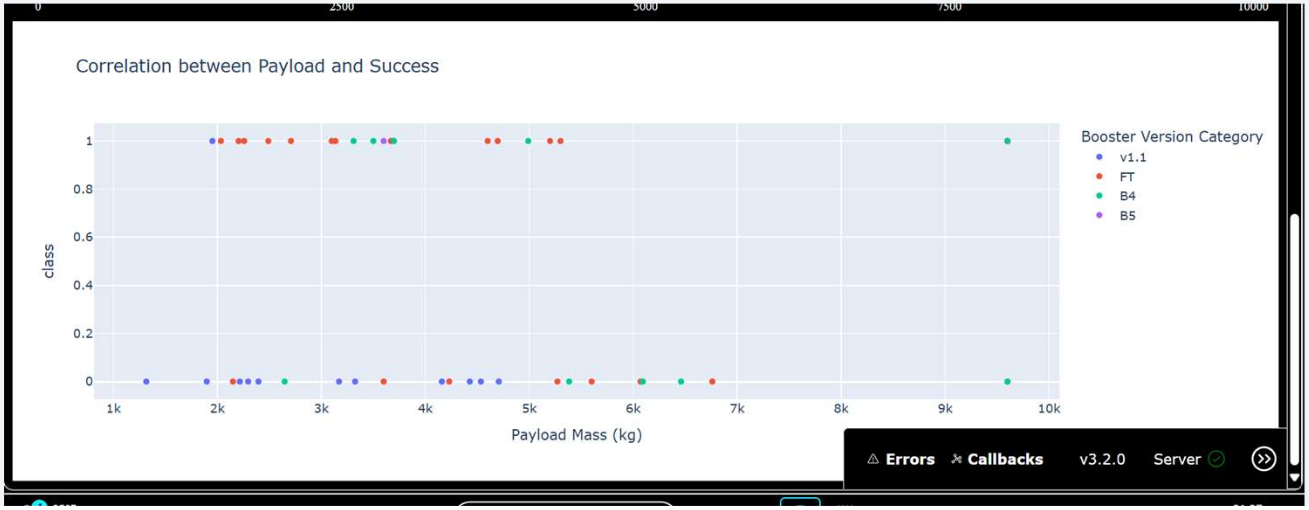
Dashboard Highest Success rate Launch Site

- -KSC LC39A had highest Successful Launches.
- - Also it was relatively large bumber of launches which means it was used for many launches.



Dashboard Payload Vs Launch Status

- - Most of the Payloads were bellow 10,000KG
- - Only few Payload are in the range of 16,000KG and most of them were successful.
- - Along with Payload mass Orbit type I is also influences the launch outcome.

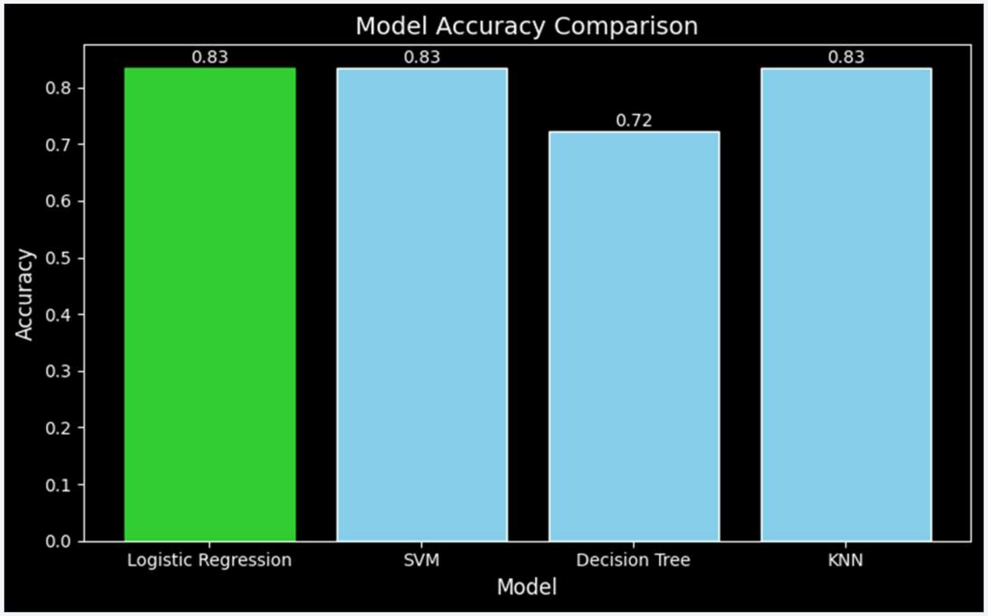


Section 5

Predictive Analysis (Classification)

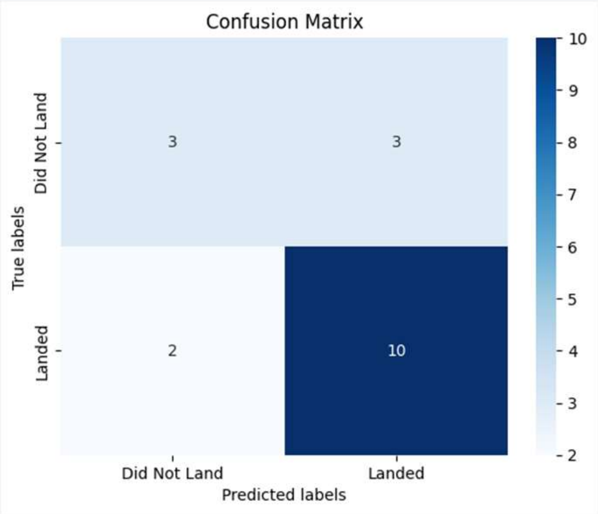
Classification Accuracy

- - Almost all models have similar accuracy.
- - Logistic Regression/Decision Tree are relatively better than other models.



Confusion Matrix

- Decision Tree had best perfuming Model.
- It's False Positive count was 3 which was a concern but it was same for all the models



Conclusions

- EDA revealed that the landing success rate has significantly improved over time, with specific Orbit Types (e.g., LEO, ISS) and the launch site KSC LC 39A showing the highest success.
- The Decision Tree Classifier model achieved the highest predictive accuracy after hyperparameter tuning via GridSearchCV, confirming that first-stage landing success is a highly predictable outcome.
- Features such as Orbit Type, Launch Site, Payload Mass etc. are important factors in Launch Success.
- Competitors of SpaceX can infer the key parameters for success and in corporate them in their launch attempts.

Appendix

- Lab 1: Collecting the data through API
- https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
- https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/jupyter-labs-webscraping.ipynb
- Lab 2: Data Wrangling
- https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
- Assignment: SQL Notebook for Peer Assignment
- https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- Assignment: Exploring and Preparing Data
- https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/edadataviz.ipynb
- Hands-on Lab: Interactive Visual Analytics with Folium
- https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/Copy%20of%20WithAnswers_lab_jupyter_launch_site_location%20.ipynb
- Hands on Lab: Complete the Machine Learning Prediction lab
- https://github.com/drythetowel/IBM_DataScience_Capstone_Project/blob/main/Updated-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!