IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Amine Saad
05/02/2024

Outline

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

Executive Summary

- Summary of the methodologies used:

  Data Collection through web scraping and SpaceX API.

  Data Wrangling through cleaning and refining collected data; converting data into a comprehensible information.

  Exploratory Data Analysis through converting the cleaned data into visualized data for better insights.

  Machine Learning Prediction through finding the best model and the best scores to finalize findings.

- Summary of the results:

  Data Collection was achievable through the mentioned sources.

  Exploratory Data Analysis through converting the cleaned data into visualized data for better insights.

  Machine Learning Prediction highlighted the best model and the optimal scores for a successful launch.

Introduction

Space exploration has always been at the forefront of human achievement, driving innovation and pushing the boundaries of what is possible. In this era, SpaceY, a pioneering aerospace company, is at the vanguard of space exploration, committed to making the cosmos accessible to all. As part of our ongoing efforts to advance space travel and ensure the safety and efficiency of our missions, we present this comprehensive documentation on the data science project conducted in collaboration with SpaceY.

This project, titled "SpaceY Rocket Launch Data Analysis," delves into the realm of data-driven decision-making for rocket launches. Launching a rocket into space is a complex and high-stakes endeavor that requires meticulous planning, monitoring, and analysis. At SpaceY, we believe that harnessing the power of data can significantly enhance the success and safety of our missions.

In the fiercely competitive landscape of the modern space industry, achieving a successful rocket launch is not only a matter of scientific achievement but also a critical business imperative. SpaceY is driven by a core mission to lead the way in space exploration, and one of our paramount challenges is to outpace our competitors by consistently achieving successful rocket launches.

Section 1

Methodology

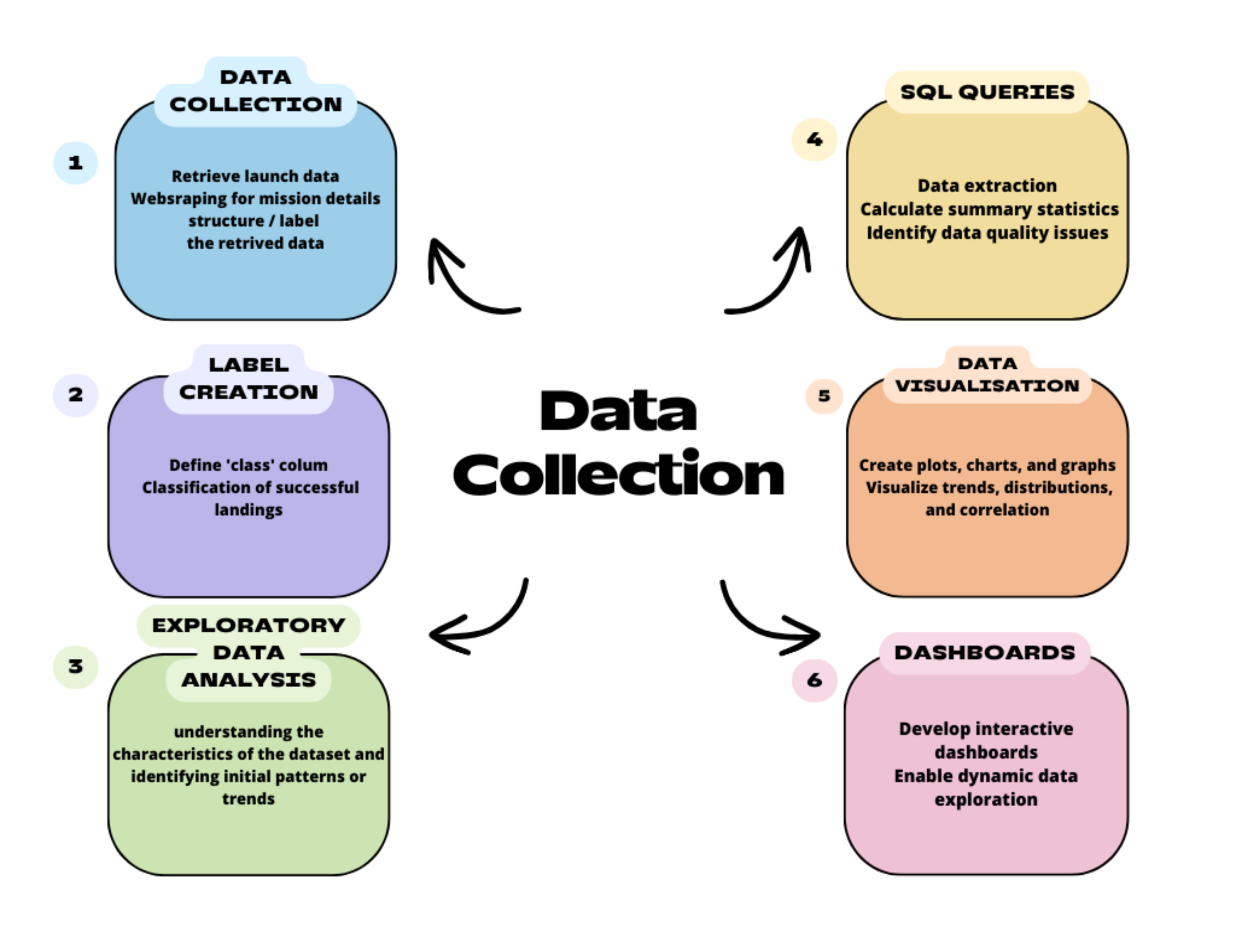
Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

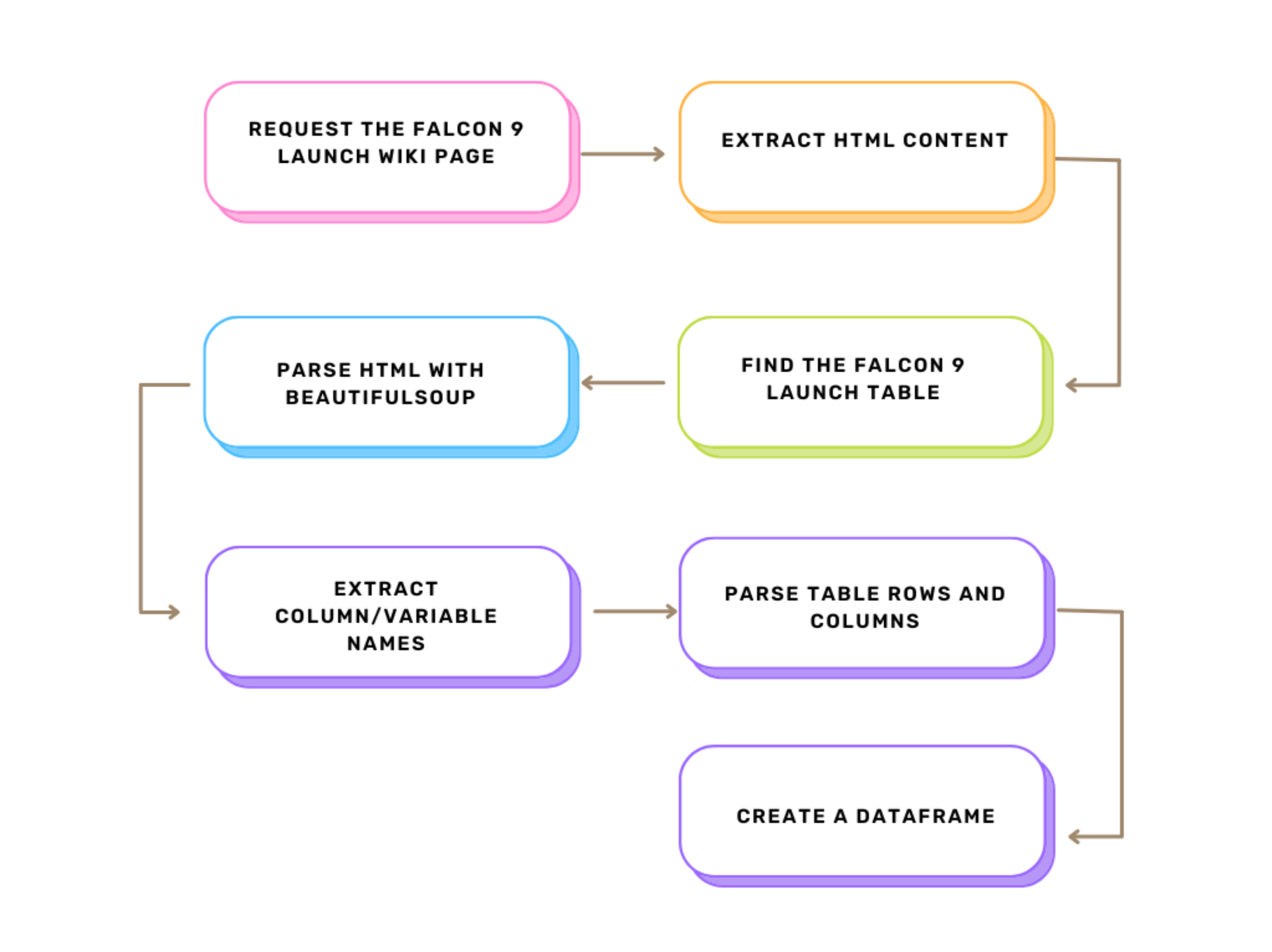
- Flowchart of Data Collection Process

- GitHub URL of the completed SpaceX API calls



Data Collection - Scraping

- Request the Falcon 9 Launch Wiki Page:
  Use requests.get(url) to fetch the page HTML.
- Extract HTML Content:
  Store the response from the website in html_content.
- Parse HTML with BeautifulSoup:
  Initialize BeautifulSoup with html_content.
- Find the Falcon 9 Launch Table:
  Use BeautifulSoup to find the table with class 'wikitable'.
- Extract Column/Variable Names:
  Loop through <th> tags in the table header.
- Parse Table Rows and Columns:
  Loop through each row (<tr>) and each cell (<td> or <th>) to extract data.
- Create a DataFrame:
  Use pd.DataFrame(rows_list) with the list of row data and column headers.



Data Wrangling

- After loading the dataset, we calculated the number of launches at each site, showing how active or preferred each launch site is. We then analyzed the orbit types, providing insights into the most commonly targeted orbits for missions. By examining the mission outcomes for each orbit, we gained an understanding of orbit-specific mission success rates or challenges. Lastly, we created a simplified landing outcome label, which could help in quickly assessing the success rate of landings.

Step 1: Load the Dataset

First, we'll import pandas and load the dataset from the provided URL.

Step 2: Calculate the Number of Launches on Each Site

Step 3: Calculate the Number and Occurrence of Each Orbit

Step 4: Calculate the Number and Occurrence of Mission Outcome of the Orbits

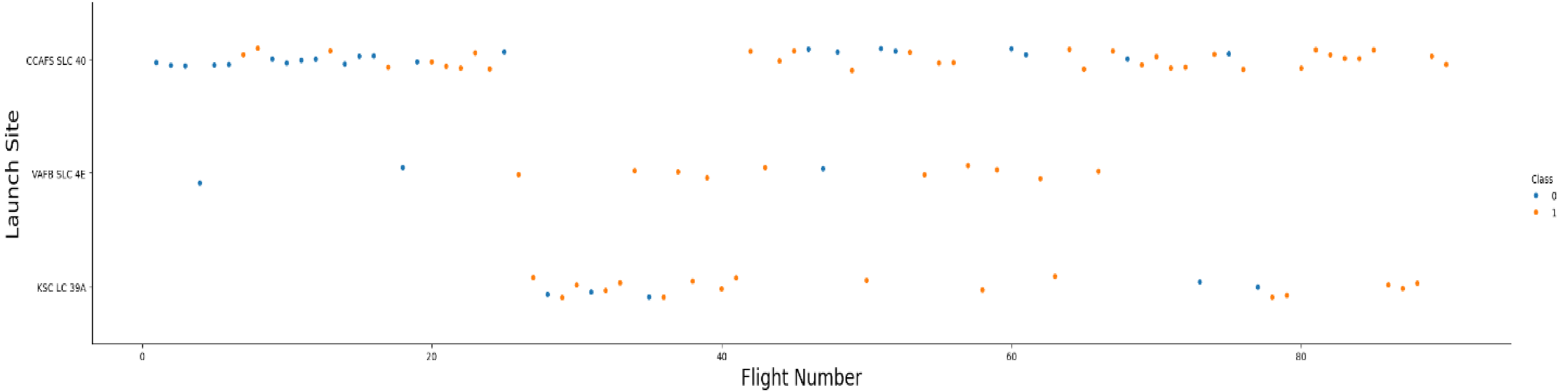
Step 5: Create a Landing Outcome Label from the Outcome Column

Assuming the Outcome column contains information about landing success or failure, we'll extract and label this.

EDA with Data Visualization

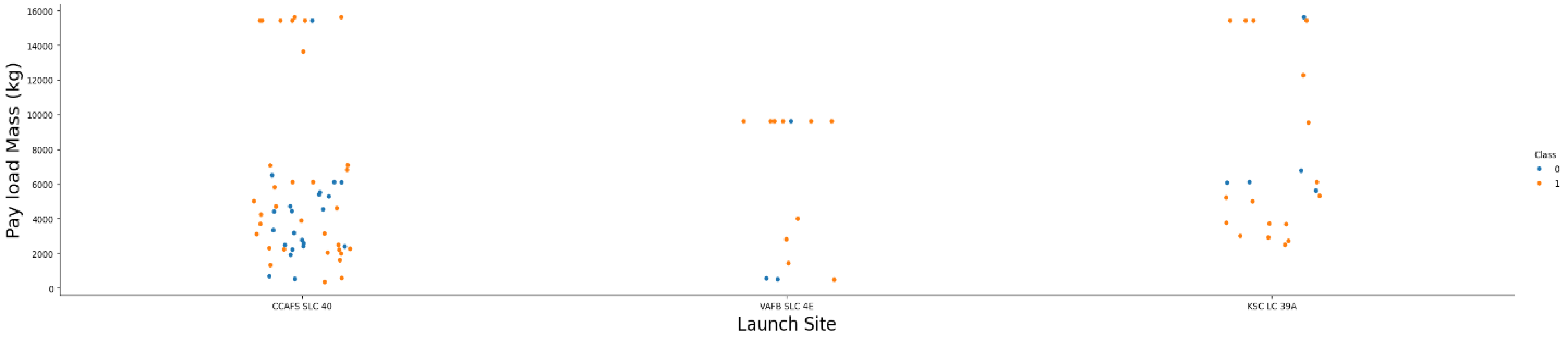
- Exploratory Data Analysis with Seaborn
  Several visuals depicting the certain variable related to launch
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

Flight Number vs Launch Site



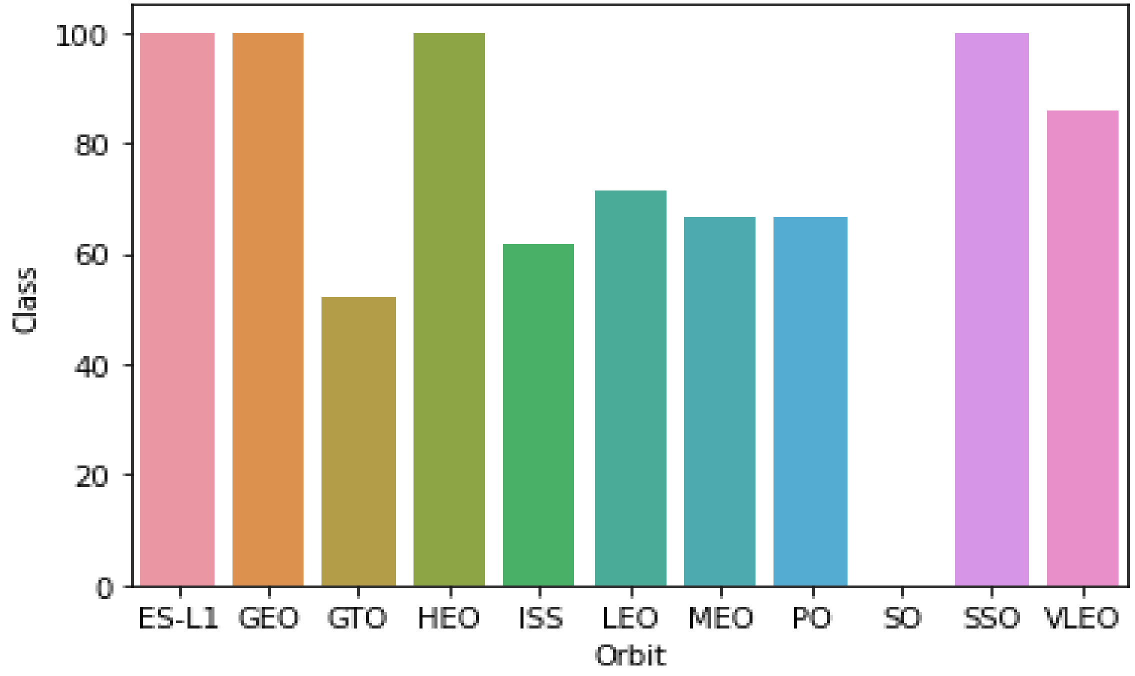
*Around the 20 flight mark CCAFS start witnessing some successful launches
*VAFB and KSC LC only confirms the continuation of the successful launches but with a lesser number
*Around flight 80 you can notice a stable success in the main site CCAFS

The Relationship Between Payload and Launch Site



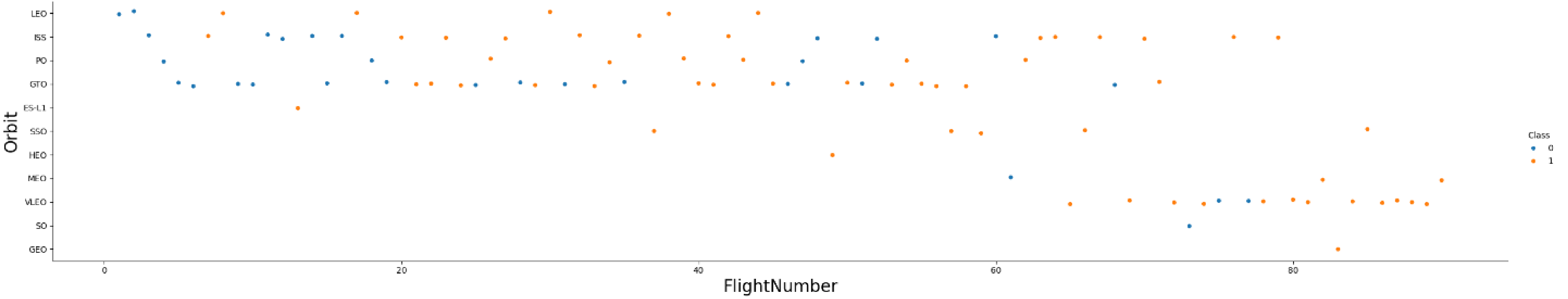
The scatter plot suggests that the CCAFS SLC 40 site has the highest launch frequency across a broad range of payload masses. VAFB SLC 4E shows fewer and generally lighter payloads, while KSC LC 39A is capable of launching heavy payloads.

The Relationship Between Success Rate of Each Orbit Type



The bar chart presents the number of successful missions for various orbit types. ES-L1, GTO, and VLEO orbits have the highest number of successes, indicating frequent and potentially reliable operations to these destinations. Orbits like HEO, ISS, LEO, and MEO show moderate success counts, suggesting regular use but with fewer missions compared to the highest ones. Orbits such as PO and SSO have lower success counts, which could imply they are less commonly targeted or possibly more challenging for missions.

the relationship between FlightNumber and Orbit type



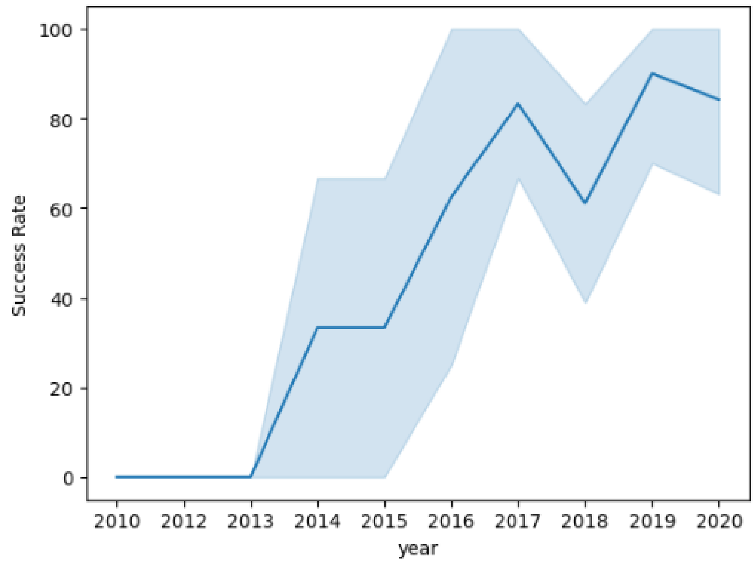
Missions span a wide range of orbit types, with LEO (Low Earth Orbit) being the most frequently targeted, as indicated by the density of data points across flight numbers.

There are distinct clusters of flight numbers for certain orbit types (e.g., GTO, Geostationary Transfer Orbit), which may suggest specific periods of focused missions to those orbits.

The two classes are present across almost all orbit types and flight numbers, indicating a mix of outcomes that are not exclusive to any particular orbit type or period.

Yearly launch Success Rate Trend

<AxesSubplot:xlabel='year', ylabel='Success Rate'>



This suggests significant improvements in launch reliability over the first half of the decade, possibly due to advancements in technology, increased experience, and improved processes. The high success rates towards the end of the decade might reflect a matured space industry with consistent performance. The shaded area indicates the confidence interval for the success rate, which seems to narrow over time, implying that the estimates for the success rate become more precise or consistent as time progresses.

EDA with SQL

- Exploratory Data Analysis with SQL with SQL DB2 in Python
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Launch Sites Names

- All the launch sites from the mission

```
In [6]:  %sql select DISTINCT LAUNCH_SITE from SPACEXDATASET

 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e3
Done.
```

Out[6]:

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

CCAFS LC-40 and CCAFS SLC-40 seem like an upgrade of the same launch site

Launch Sites Starting with CCA

Despite the mission outcome being marked as 'Success' for all listed flights, the 'landing_outcome' shows mixed results with some attempts resulting in 'Failure (parachute)' and others with 'No attempt' to land.
Early Falcon 9 Missions: The booster versions listed (F9 v1.0) suggest these are early missions for SpaceX's Falcon 9 vehicle.

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

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

The Total Payload Mass Carried By Boosters Launched By NASA (CRS)

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.a
Done.
```

SUM
45596

Average Payload Mass Carried By Booster Version F9 V1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACEXDATASET where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.c
Done.
```

average
2534

The Date When The First Successful Landing Outcome In Ground Pad Was Acheived.

```
%sql select min(date) as Date from SPACEXDATASET where mission_outcome like 'Success'
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.data
Done.
```

DATE
2010-06-04

The Total Number Of Successful And Failure Mission Outcomes

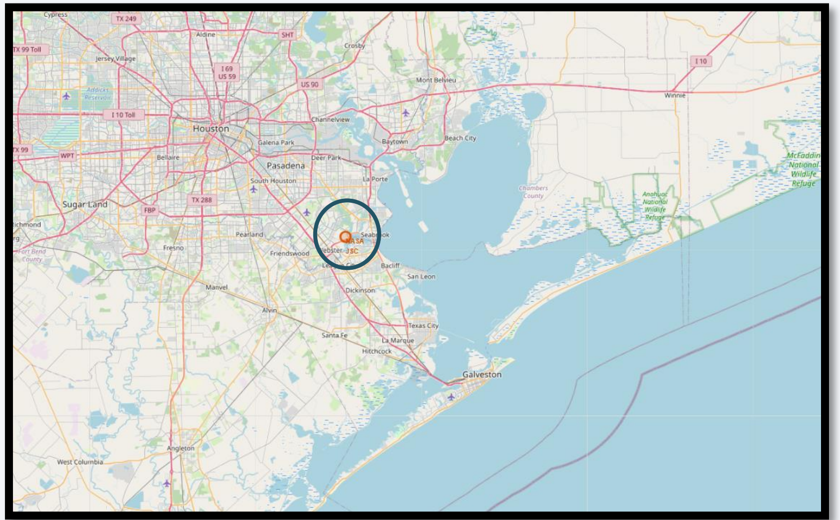
```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXDATASET GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLU
Done.
```

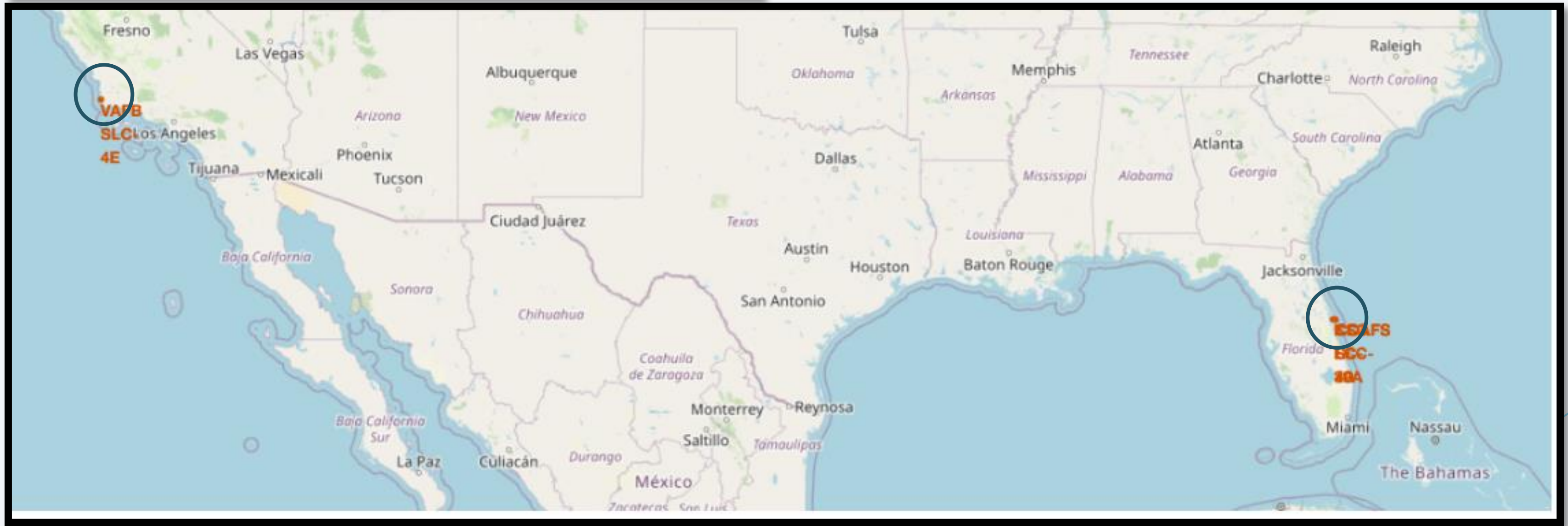
mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Interactive Map with Folium

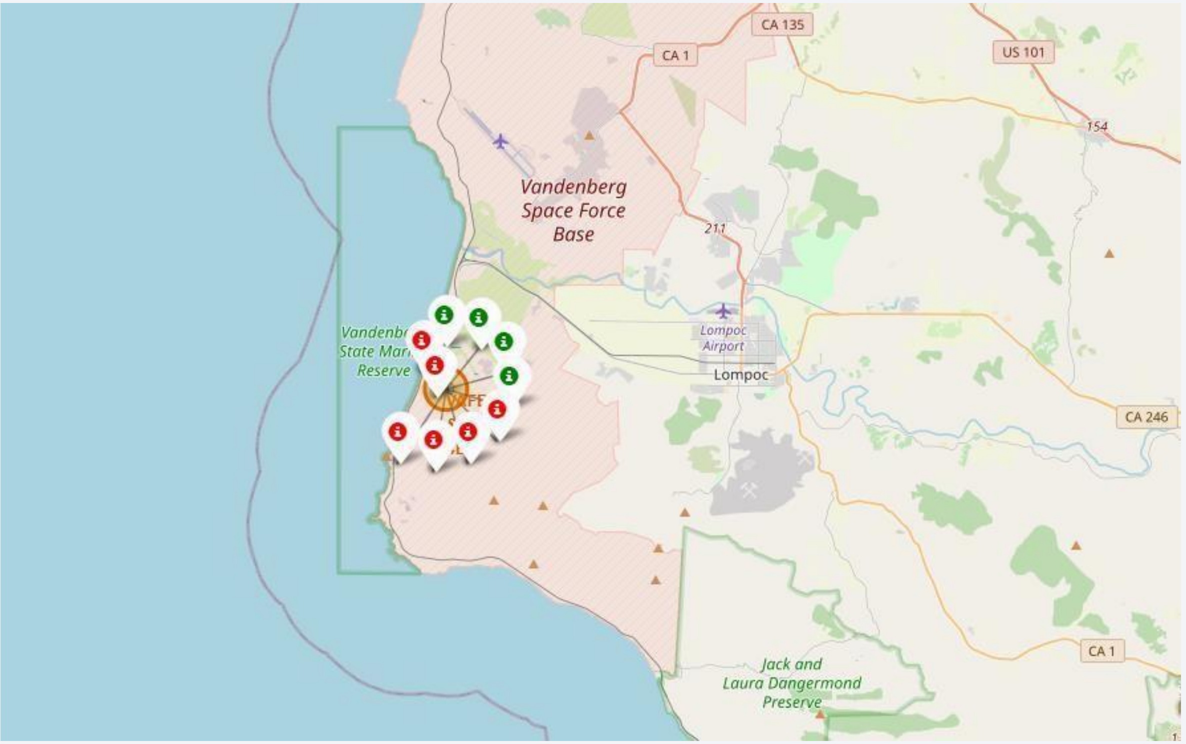
Launch Sites Locations



All launch locations are near water surfaces, mainly oceans to avoid crashing in urban areas.
There is strategic spread around the US perhaps to test have different environmental conditions.

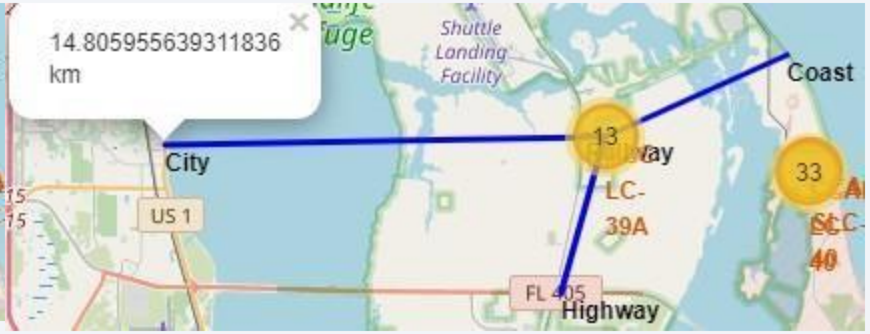


Failed / Successful Landings



Landing Zone Activity: The map shows a concentration of landing attempts within a specific area of the Vandenberg Space Force Base, which indicates it is a designated landing zone for space missions.

Failed / Successful Landings

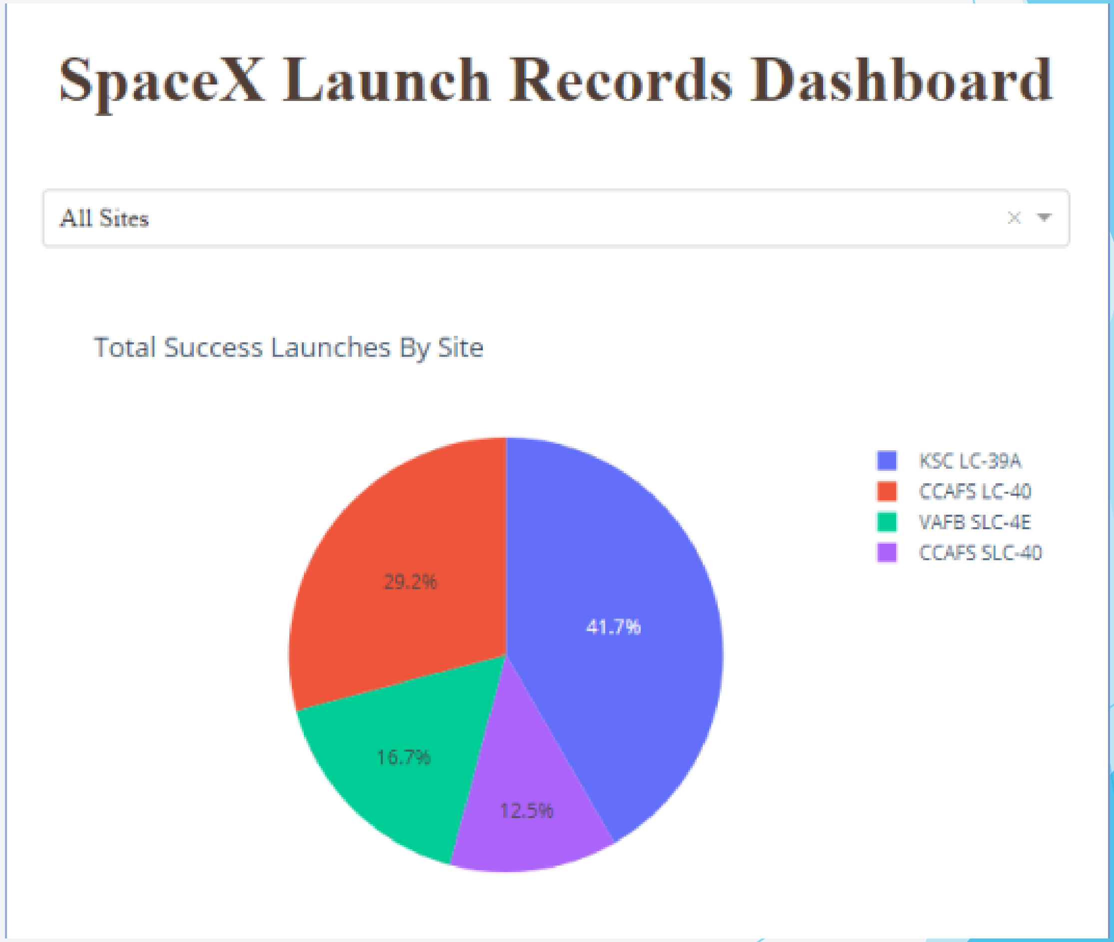


launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

Build a Dashboard with Plotly Dash

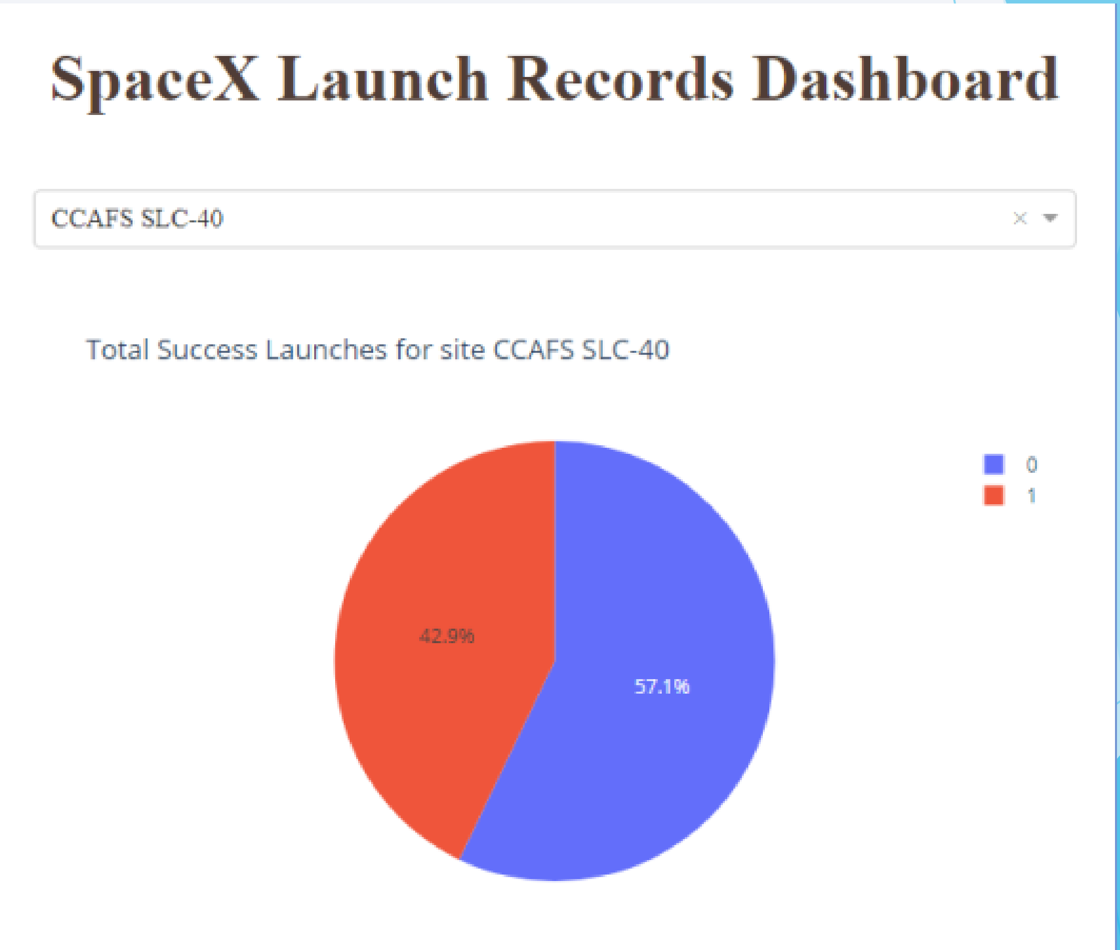
Launch Success Count for All Sites

- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A.



Launch Site with Highest Launch Success Ratio

- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%).

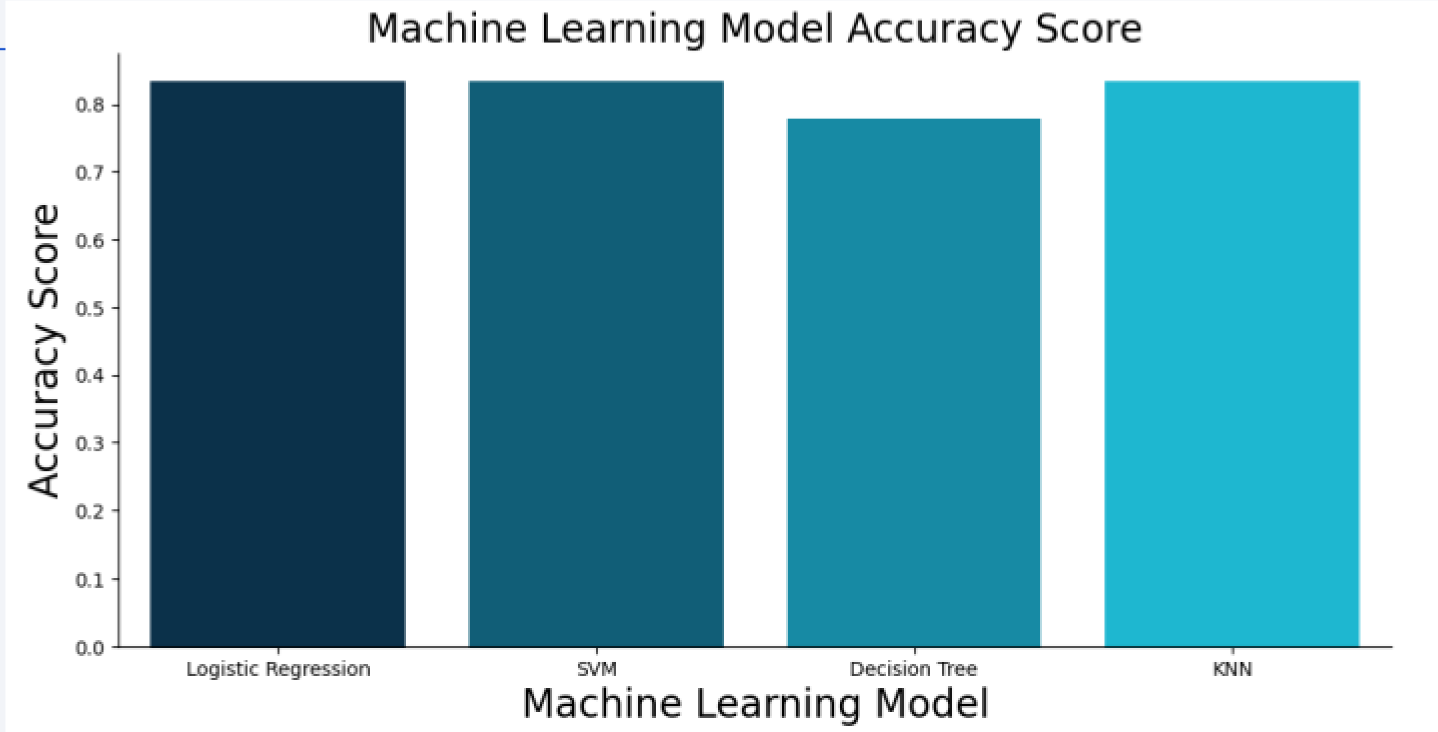


Categories Comparison

- Larger dots, which indicate a higher number of launches, are scattered across various payload masses, suggesting frequent launches irrespective of payload weight.
- In the payload mass range of 0-6,000 kg, most launches resulted in successful landings (class 1). However, there are two notable instances where the payload mass was zero kg, and the landings were unsuccessful (class 0).
- There doesn't seem to be a clear pattern linking booster version categories to either success or failure of landings.
- The plot suggests that payload mass does not directly determine the success of the landing, as successes and failures are present across a wide range of payload masses.

Predictive Analysis (Classification)

Categories Comparison



- Larger dots, which indicate a higher number of launches, are scattered across various payload masses, suggesting frequent launches irrespective of payload weight.
- In the payload mass range of 0-6,000 kg, most launches resulted in successful landings (class 1). However, there are two notable instances where the payload mass was zero kg, and the landings were unsuccessful (class 0).
- There doesn't seem to be a clear pattern linking booster version categories to either success or failure of landings.
- The plot suggests that payload mass does not directly determine the success of the landing, as successes and failures are present across a wide range of payload masses.

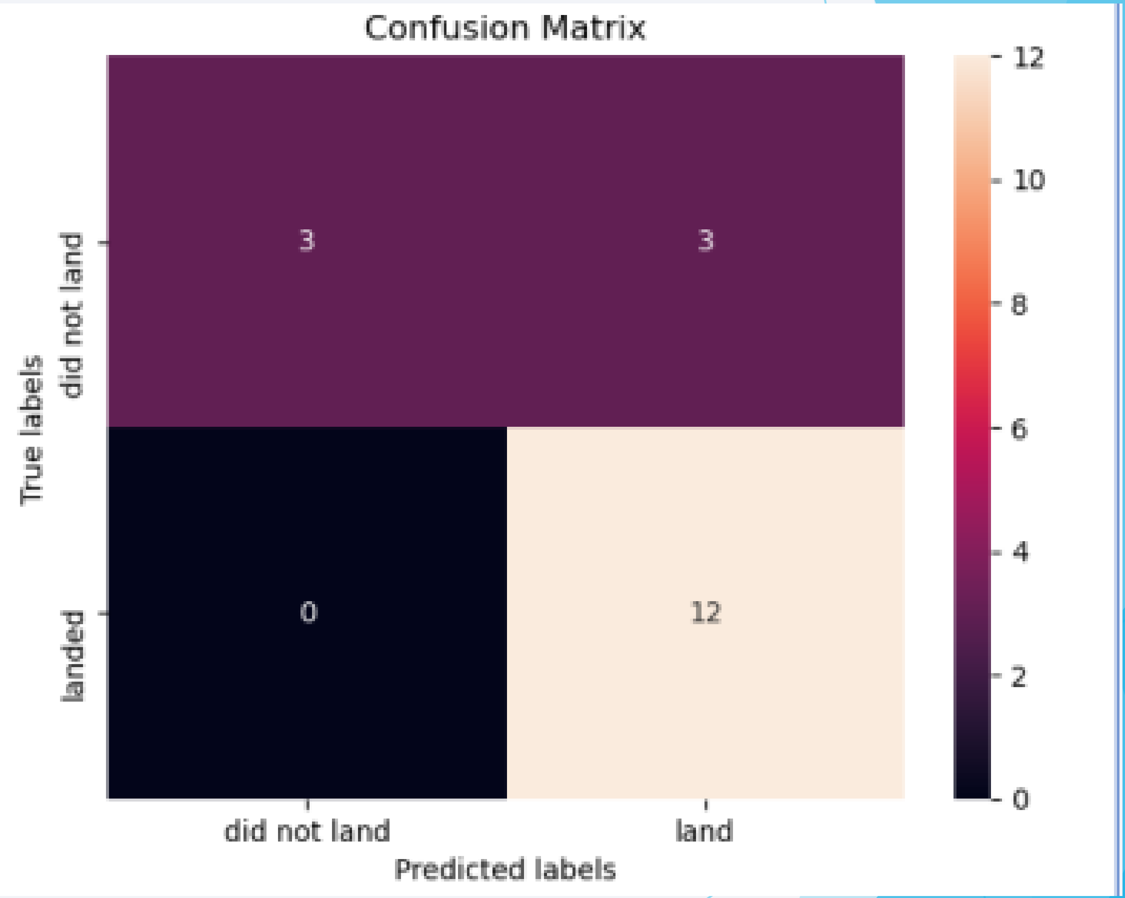
Confusion Matrix

- The matrix shows that there were 3 instances where the model correctly predicted that the spacecraft would not land ("did not land" is true, and the model predicted "did not land"). This is a True Negative (TN).

- There were 12 instances where the model correctly predicted that the spacecraft would land ("landed" is true, and the model predicted "land"). This is a True Positive (TP).

- There were 3 instances where the model incorrectly predicted that the spacecraft would land (the spacecraft "did not land", but the model predicted "land"). This is a False Positive (FP).

- There were no instances where the model incorrectly predicted that the spacecraft would not land (the spacecraft "landed", but the model predicted "did not land"). This is a False Negative (FN).



Conclusions

- The project successfully leveraged data science methodologies to create a predictive model with high accuracy for the binary classification task of predicting landing outcomes.
- The model showed strength in identifying successful landings with few misclassifications, suggesting that the selected features and model were well-suited for the task.
- The exploratory analysis provided valuable insights into operational patterns and potential factors influencing mission outcomes, which were crucial for feature selection in the modeling phase.
- Future enhancements could involve fine-tuning the model to reduce the false positives and incorporating additional data as it becomes available.
- The project demonstrates the potential of machine learning in enhancing decision-making processes in aerospace endeavors by providing predictive insights that could potentially increase the success rates of future missions.
- It's a testament to the power of data science in solving complex real-world problems and advancing the field of space exploration.

Appendix

- **Instructors:**

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

- GitHub Project Full: https://github.com/amine711/Amine-Sd-Capstone-Project-IBM-DS

Thank you!