IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Paul Yu>
<27.3.2024>

Outline

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

Executive Summary

- Using data collected from both the public SpaceX API and the SpaceX Wikipedia page, I constructed a dataset with a 'class' column to label successful landings. I delved into the data using SQL queries, visualizations, Folium maps, and dashboards. Relevant columns were extracted as features, and categorical variables were converted into binary using one-hot encoding. Standardization was applied to the data, followed by employing GridSearchCV to fine-tune parameters for the machine learning models.

- Subsequently, four machine learning models were developed: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. Remarkably, all models exhibited similar performances with an accuracy rate hovering around 80%. It's notable that all models tended to over-predict successful landings. This suggests that additional data would likely enhance model determination and accuracy, thereby yielding more reliable predictions.

Introduction

Background:

- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

- Problem:

- Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

Section 1

Methodology

Methodology

The data collection process involved amalgamating information from both the publicly available SpaceX API and the SpaceX Wikipedia page. Subsequently, extensive data wrangling was conducted to clean and prepare the dataset for analysis.

The data was then classified, distinguishing successful landings as true positives and all other outcomes as unsuccessful. Following this, exploratory data analysis (EDA) was undertaken utilizing visualization techniques and SQL queries to uncover insights and patterns within the dataset.

Interactive visual analytics were further conducted using tools such as Folium for geospatial analysis and Plotly Dash for dynamic dashboards, enhancing the exploration and presentation of the data.

Moving forward, predictive analysis was performed utilizing classification models to forecast successful landings. These models were fine-tuned employing GridSearchCV to optimize their parameters for improved performance and accuracy.

Data Collection

- The data collection process involved a dual approach, comprising API requests from SpaceX's public API and web scraping data from a table within SpaceX's Wikipedia entry.

- The subsequent slide will illustrate the flowchart detailing the data collection process from the API, while the following slide will showcase the flowchart depicting the data collection process from web scraping.

- Columns extracted from the SpaceX API data include FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

- On the other hand, the Wikipedia web scrape yielded columns such as Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Booster Version, Booster landing, Date, and Time.

Data Collection – SpaceX API

GitHub URL

https://github.com/Sagnik010/IBM DataScience/blob/main/Data%20Collection%20Api%20.ipynb

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

GitHub URL

https://github.com/Sagnik010/IBMDataScience/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Place your flowchart of web scraping here

Data Wrangling

- To generate a training label based on landing outcomes, we'll focus on two components within the 'Outcome' column: 'Mission Outcome' and 'Landing Location'. We'll create a new column named 'class' to represent this label, assigning a value of 1 if 'Mission Outcome' is deemed successful and 0 otherwise. Here's how we'll map the values:

- For 'Mission Outcome' values of 'True' with corresponding 'Landing Location' of 'ASDS', 'RTLS', or 'Ocean', we'll set 'class' to 1.
- For all other combinations, including 'None' or 'False' for 'Mission Outcome' and various 'Landing Location' values, 'class' will be set to 0.

EDA with Data Visualization

- Exploratory Data Analysis (EDA) was conducted on several variables including Flight Number, Payload Mass, Launch Site, Orbit, Class (indicating landing success), and Year. The aim was to investigate potential relationships among these variables to determine their suitability for training the machine learning model.
- Various plots were utilized for this analysis:
- Scatter plots of Flight Number versus Payload Mass, Flight Number versus Launch Site, and Payload Mass versus Launch Site were employed to observe any discernible patterns or correlations between these variables.
- Bar plots were used to visualize the distribution of orbits and their corresponding success rates, providing insights into the relationship between orbit types and successful landings.
- Line charts were utilized to depict the yearly trend of successful landings, indicating any patterns or trends over time.
- By examining these plots, we aimed to identify any significant relationships or trends that could potentially inform the training of the machine learning model.

EDA with SQL

- The dataset was imported into an IBM DB2 Database, facilitating seamless data management and querying. Using SQL-Python integration, a series of queries were executed to gain deeper insights into the dataset.
- Specific queries were formulated to extract information regarding:
- Launch site names: To understand the distribution of launches across different sites.
- Mission outcomes: To assess the success rates of missions.
- Various payload sizes of customers: To analyze the diversity and magnitude of payloads.
- Booster versions: To identify the different versions of boosters utilized.
- Landing outcomes: To examine the success or failure of landing attempts.
- By executing these queries, a comprehensive understanding of the dataset was obtained, enabling informed decision-making and analysis.

Build an Interactive Map with Folium

- Folium maps were utilized to visually represent key aspects of the data, including launch sites, successful and unsuccessful landings, as well as proximity to significant locations such as railways, highways, coastlines, and cities.

- This approach aids in understanding the rationale behind the selection of launch site locations, providing insights into factors such as accessibility and geographical considerations. Furthermore, it allows for the visualization of successful landing occurrences in relation to their geographic context, facilitating analysis of landing success relative to location.

Build a Dashboard with Plotly Dash

- The dashboard features two interactive components: a pie chart and a scatter plot.

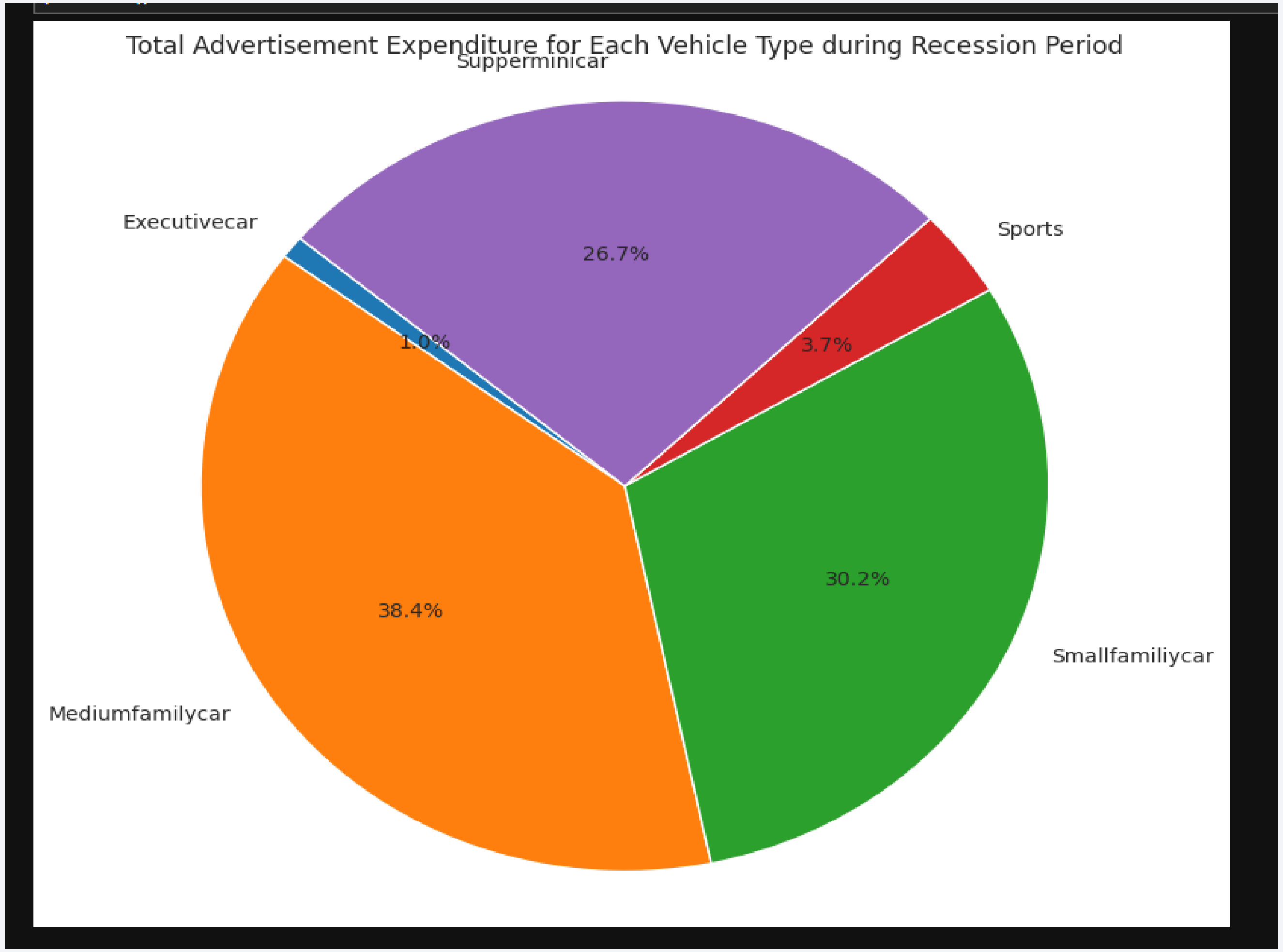
- The pie chart offers the flexibility to display the distribution of successful landings across all launch sites collectively, or it can be adjusted to showcase the success rates of individual launch sites. This visualization aids in understanding the success rates associated with different launch sites.

- On the other hand, the scatter plot allows users to select either all launch sites or an individual site. Additionally, users can adjust the payload mass using a slider ranging from 0 to 10,000 kg. This scatter plot facilitates the exploration of how success varies across launch sites, payload masses, and categories of booster versions, providing valuable insights into the relationship between these variables.

Predictive Analysis (Classification)

- Conduct exploratory data analysis (EDA) to understand the dataset and establish training labels.
- Introduce a new column, 'class', to represent the training labels.
- Standardize the dataset to ensure uniformity and comparability of features.
- Partition the dataset into training and testing subsets.
- Utilize techniques such as GridSearchCV to identify optimal hyperparameters for Support Vector Machines (SVM), Classification Trees, and Logistic Regression models.
- Evaluate the performance of each model using the test dataset.
- Determine the most effective method based on its performance with the test data.

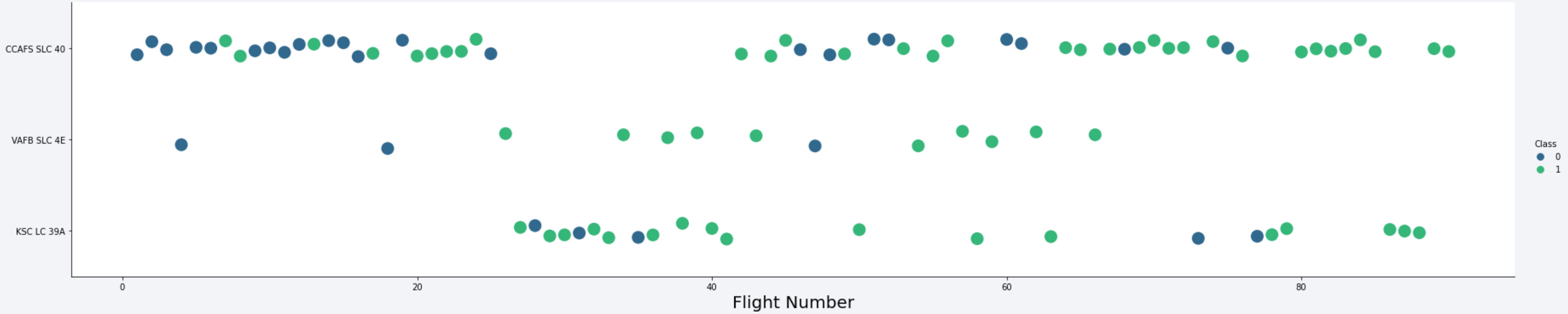
Results


Total Advertisement Expenditure for Each Vehicle Type during Recession Period
Supperminicar
26.7%
Executivecar
1.0%
Sports
3.7%
30.2%
Smallfamiliycar
38.4%
Mediumfamilycar

Section 2

Insights drawn from EDA

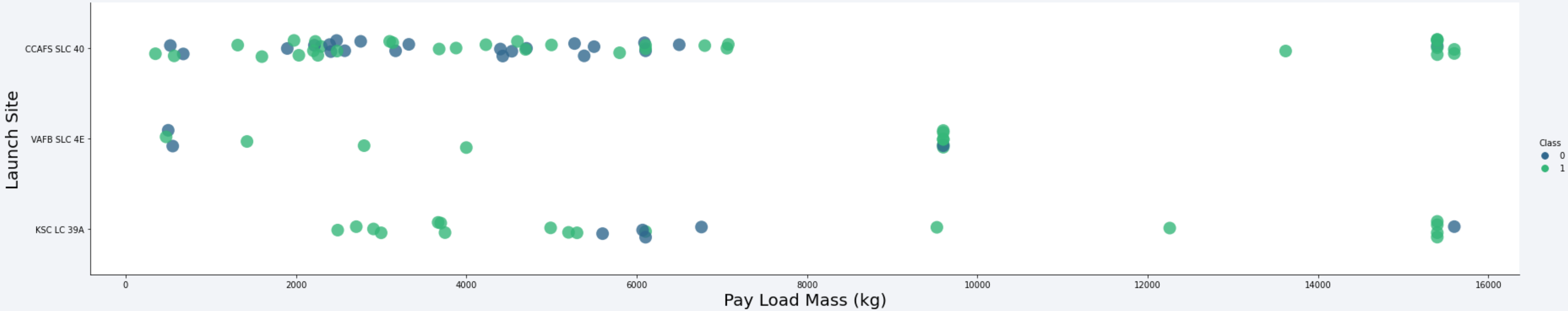
Flight Number vs. Launch Site



- The graphic illustrates successful launches in green and unsuccessful launches in purple. There appears to be a discernible trend indicating an uptick in success rates over time, as evidenced by the Flight Number. Notably, there seems to be a significant breakthrough around Flight 20, correlating with a substantial increase in success rates.

- Moreover, the analysis suggests that Cape Canaveral Air Force Station (CCAFS) serves as the primary launch site, indicated by its higher volume of launches compared to other sites.

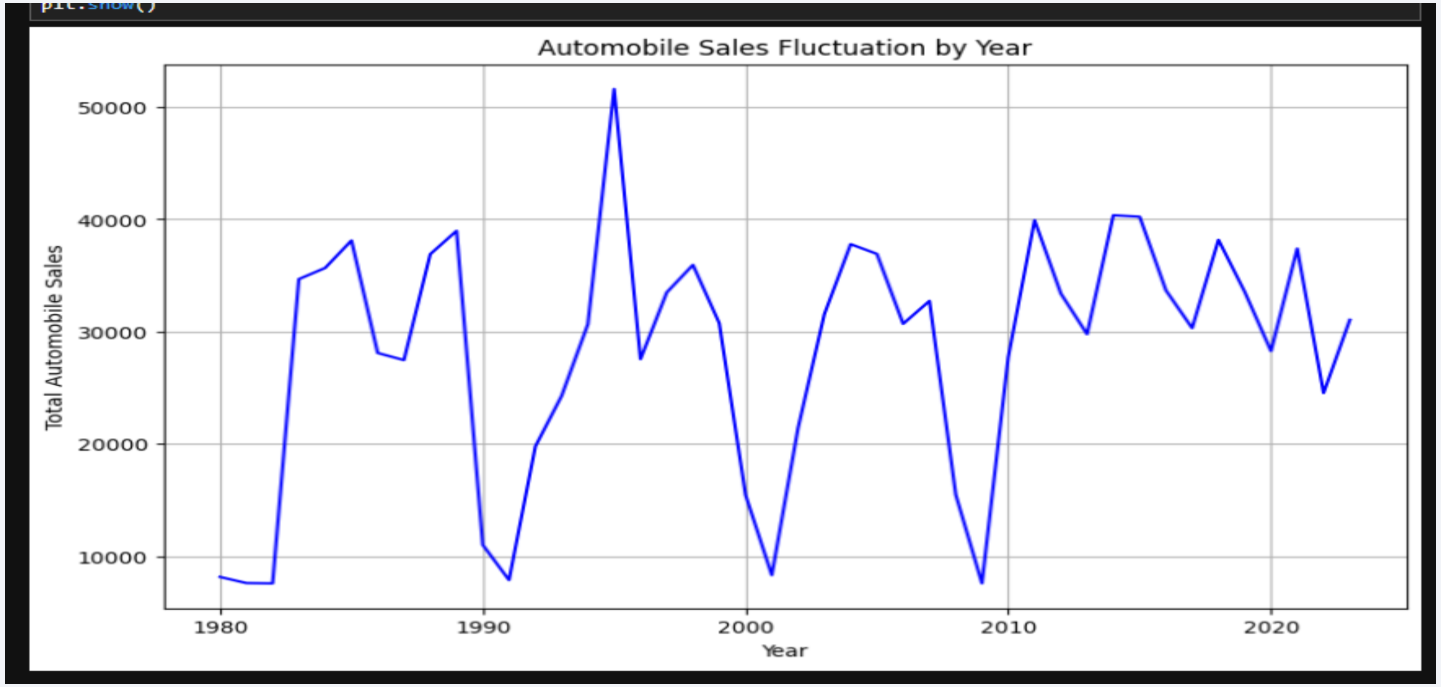
Payload vs. Launch Site



- In the visualization, successful launches are depicted in green, while unsuccessful launches are represented in purple.

- Observations suggest that the majority of payload masses fall within the range of 0-6000 kg. Additionally, it appears that different launch sites have distinct preferences or requirements regarding payload mass, as evidenced by variations in the distribution across different launch sites.

Launch Success Yearly Trend



```
plt.show()
```

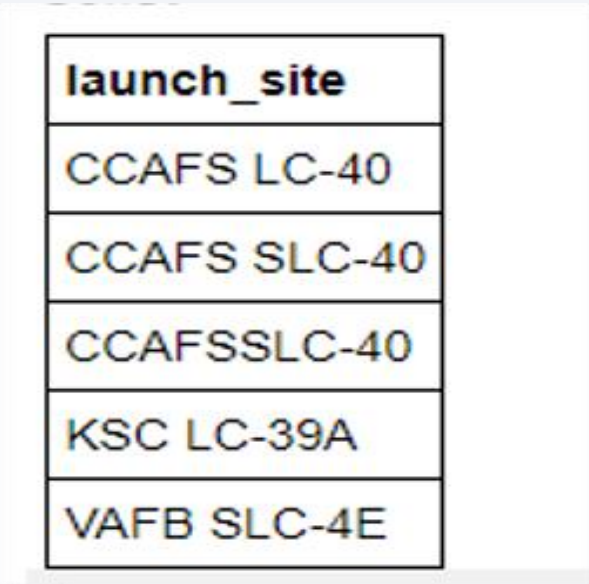
All Launch Site Names

- Retrieve the unique launch site names from the database.

- It's observed that there might be data entry errors, as "CCAFS SLC-40" and "CCAFSSLC-40" likely refer to the same launch site.

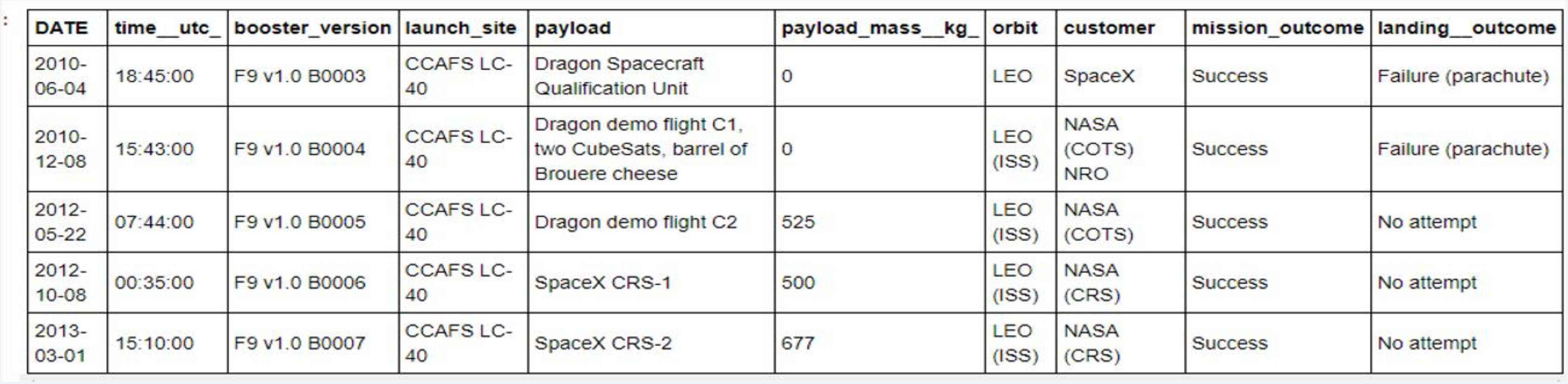
- The previous name, "CCAFS LC-40," is likely equivalent to "CCAFS SLC-40."

- Therefore, there are likely only three unique launch site values: "CCAFS SLC-40," "KSC LC-39A," and "VAFB SLC-4E." here



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

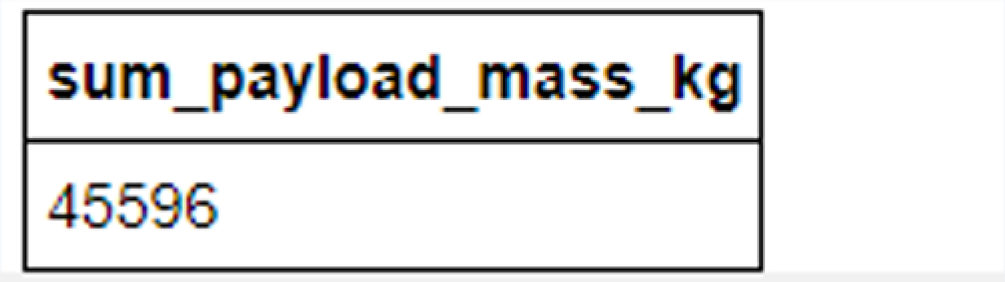
Launch Site Names Begin with 'CCA'



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

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Total Payload Mass

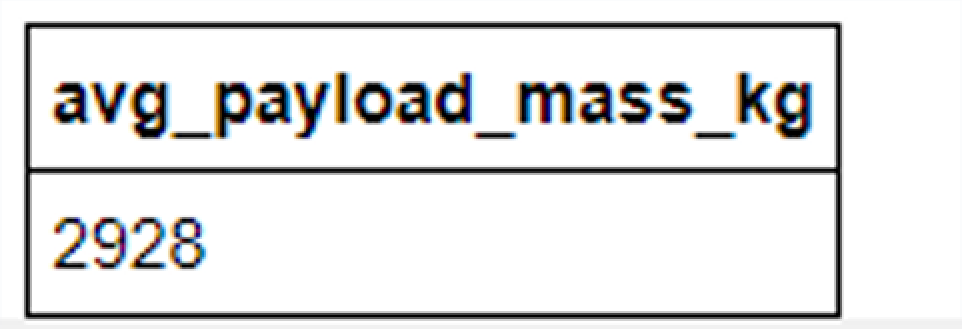


sum_payload_mass_kg
45596

- Execute the query to calculate the total payload mass, in kilograms, where NASA was the customer.

- The payloads associated with NASA typically denote missions related to the International Space Station (ISS) under the Commercial Resupply Services (CRS) program.

Average Payload Mass by F9 v1.1



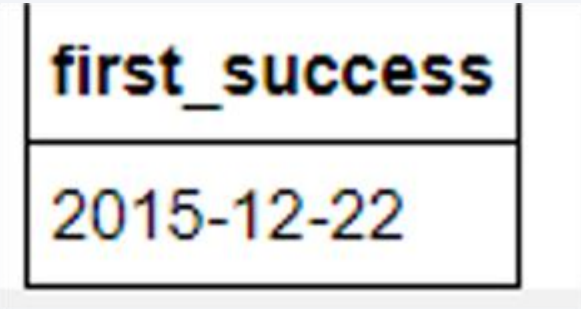
avg_payload_mass_kg
2928

- Execute the query to compute the average payload mass for launches utilizing the booster version F9 v1.1.

- The average payload mass for F9 v1.1 launches tends to fall towards the lower end of our payload mass range.

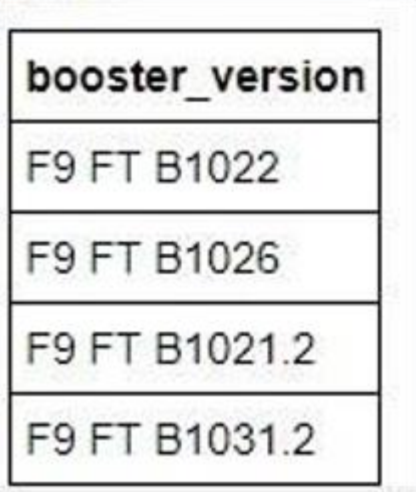
First Successful Ground Landing Date

- First Successful Ground Landing Date is 2015.12.22



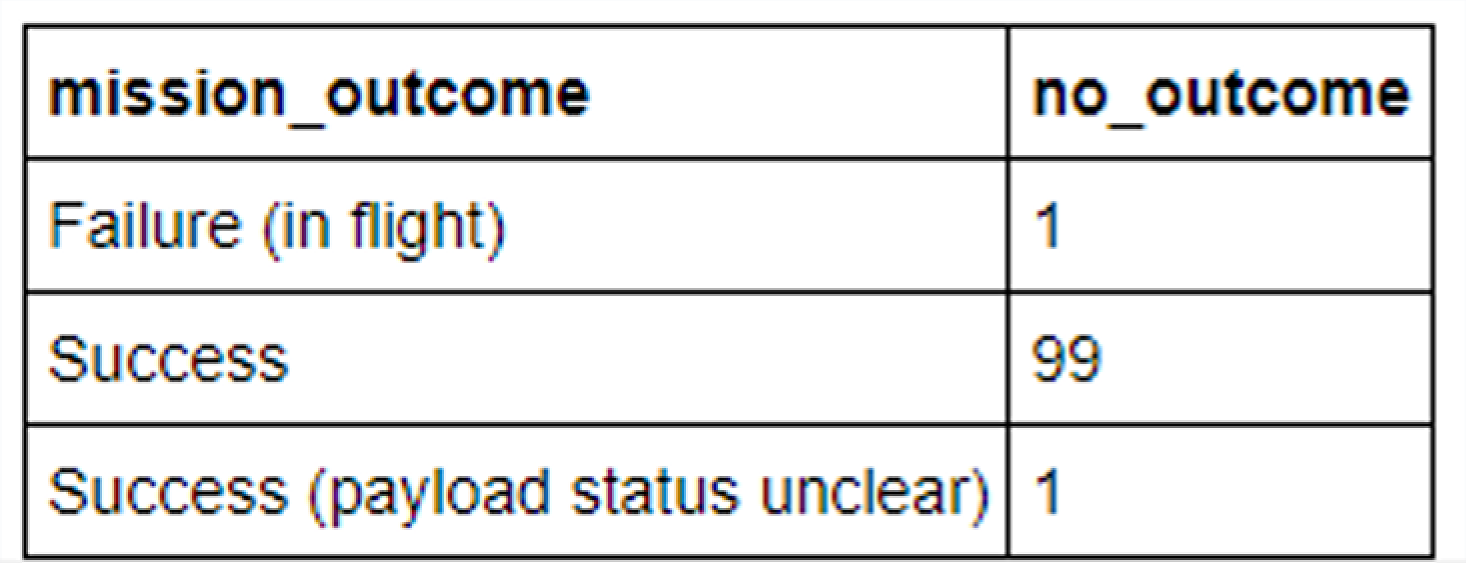
first_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000



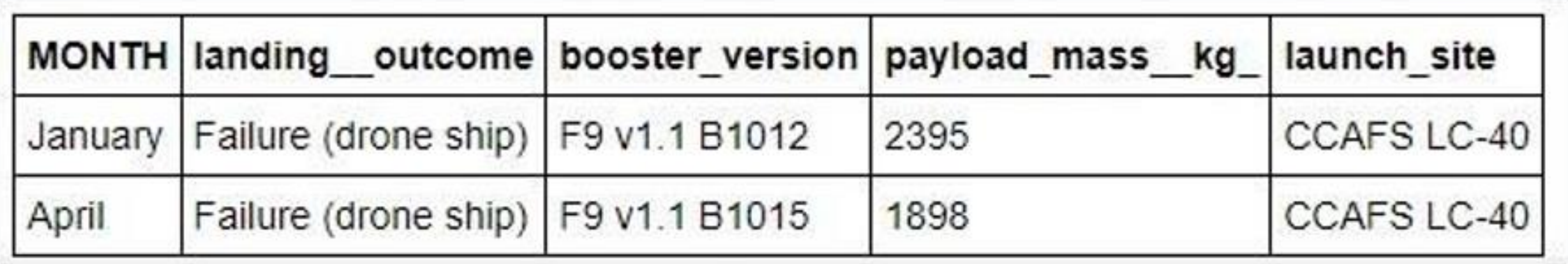
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes



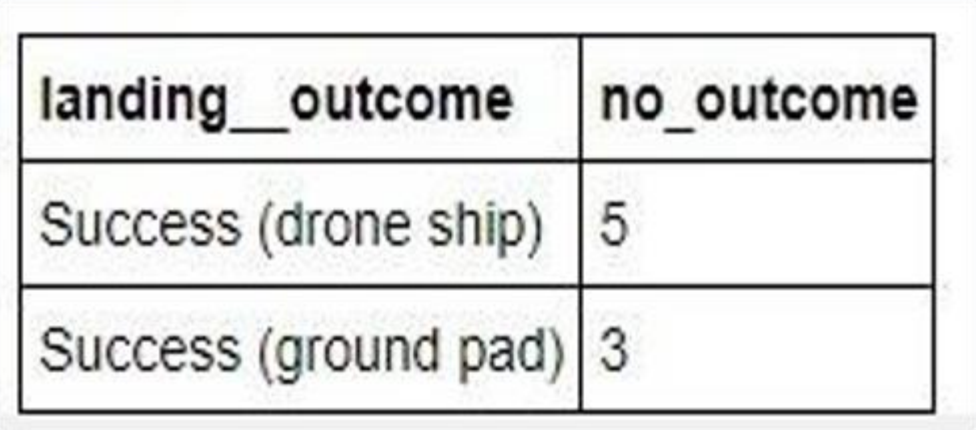
mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

2015 Launch Records



MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

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

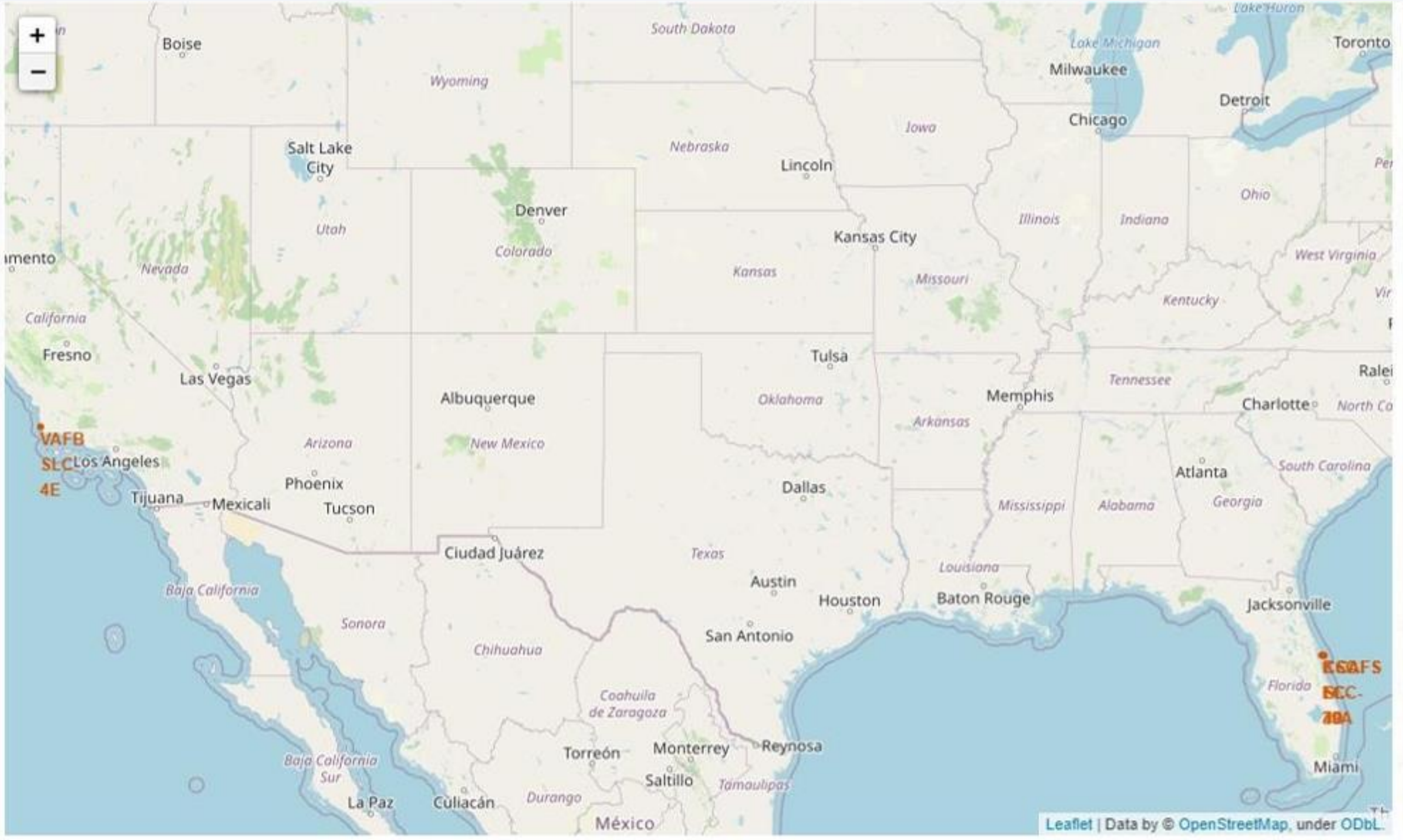


landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

Section 3

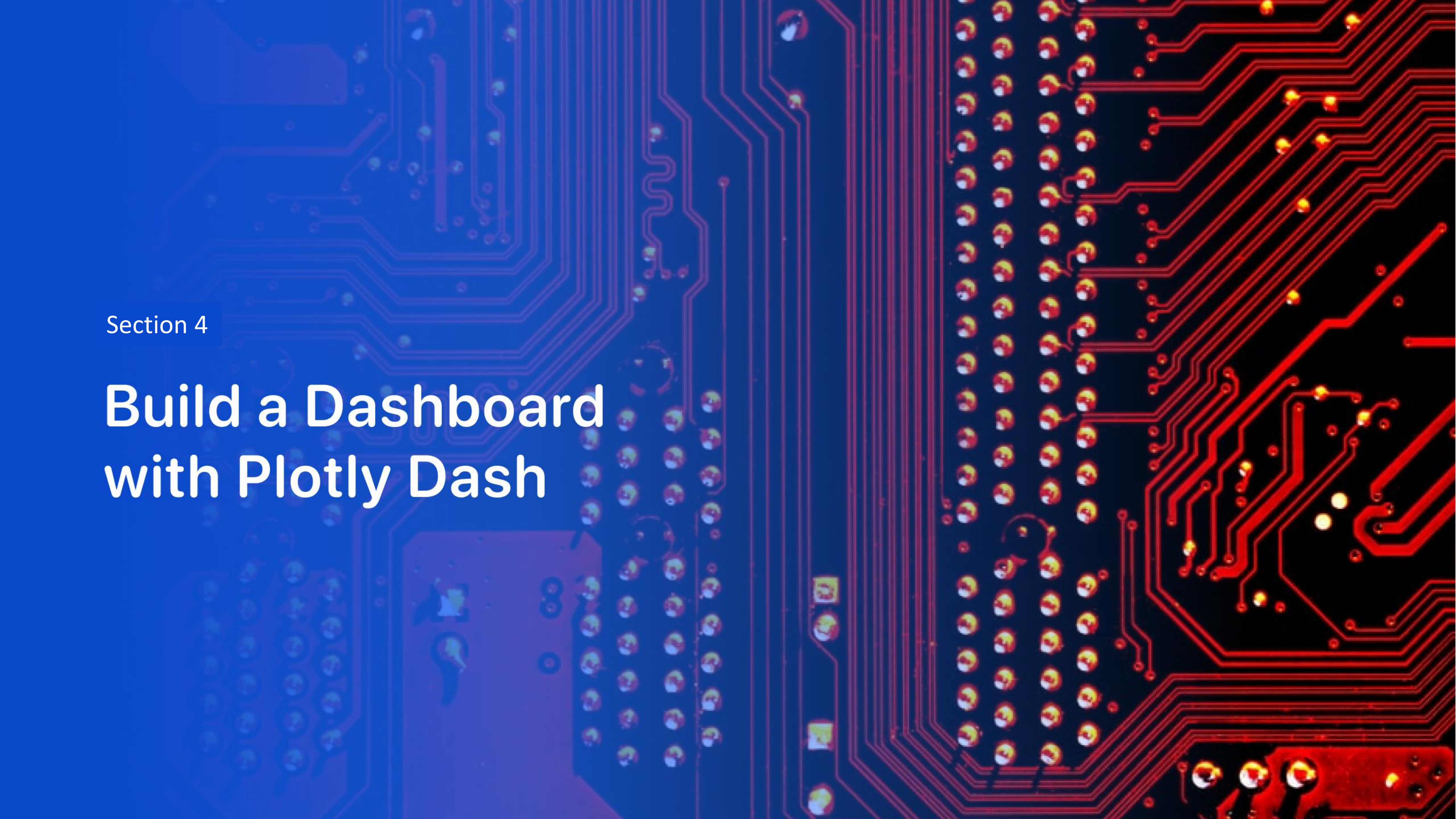
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>


+
−
VAFB
SLC-4E
CCAFS
SLC-40
LC-39A
Boise
Wyoming
South Dakota
Salt Lake City
Nebraska
Lincoln
Iowa
Milwaukee
Chicago
Detroit
Toronto
Lake Michigan
Lake Huron
Denver
Colorado
Utah
Nevada
California
Fresno
Las Vegas
Kansas
Kansas City
Missouri
Illinois
Indiana
Ohio
West Virginia
Kentucky
Tulsa
Oklahoma
Memphis
Tennessee
Arkansas
Albuquerque
New Mexico
Arizona
Los Angeles
Phoenix
Tucson
Tijuana
Mexicali
Charlotte
North Carolina
South Carolina
Atlanta
Georgia
Alabama
Mississippi
Dallas
Texas
Austin
Houston
San Antonio
Louisiana
Baton Rouge
Jacksonville
Florida
Miami
Ciudad Juárez
Baja California
Sonora
Chihuahua
Coahuila de Zaragoza
Torreón
Monterrey
Saltillo
Reynosa
Tamaulipas
Durango
Culiacán
La Paz
Baja California Sur
México
Leaflet | Data by © OpenStreetMap, under ODbL.

Section 4

Build a Dashboard with Plotly Dash



<Dashboard1>

- The distribution of successful landings across all launch sites reveals some noteworthy patterns. The old name "CCAFS LC-40" is now recognized as "CCAFS SLC-40," suggesting that successful landings at both CCAFS and KSC are attributed to the same launch site. However, it's worth noting that a significant proportion of successful landings occurred before the name change.

- In contrast, VAFB has the smallest share of successful landings. This could be attributed to a smaller sample size of launches from this site and potentially increased operational challenges associated with launching from the west coast.

<Dashboard2>

- there's a Payload range selector that currently spans from 0 to 10,000, which doesn't fully capture the maximum payload value of 15,600.

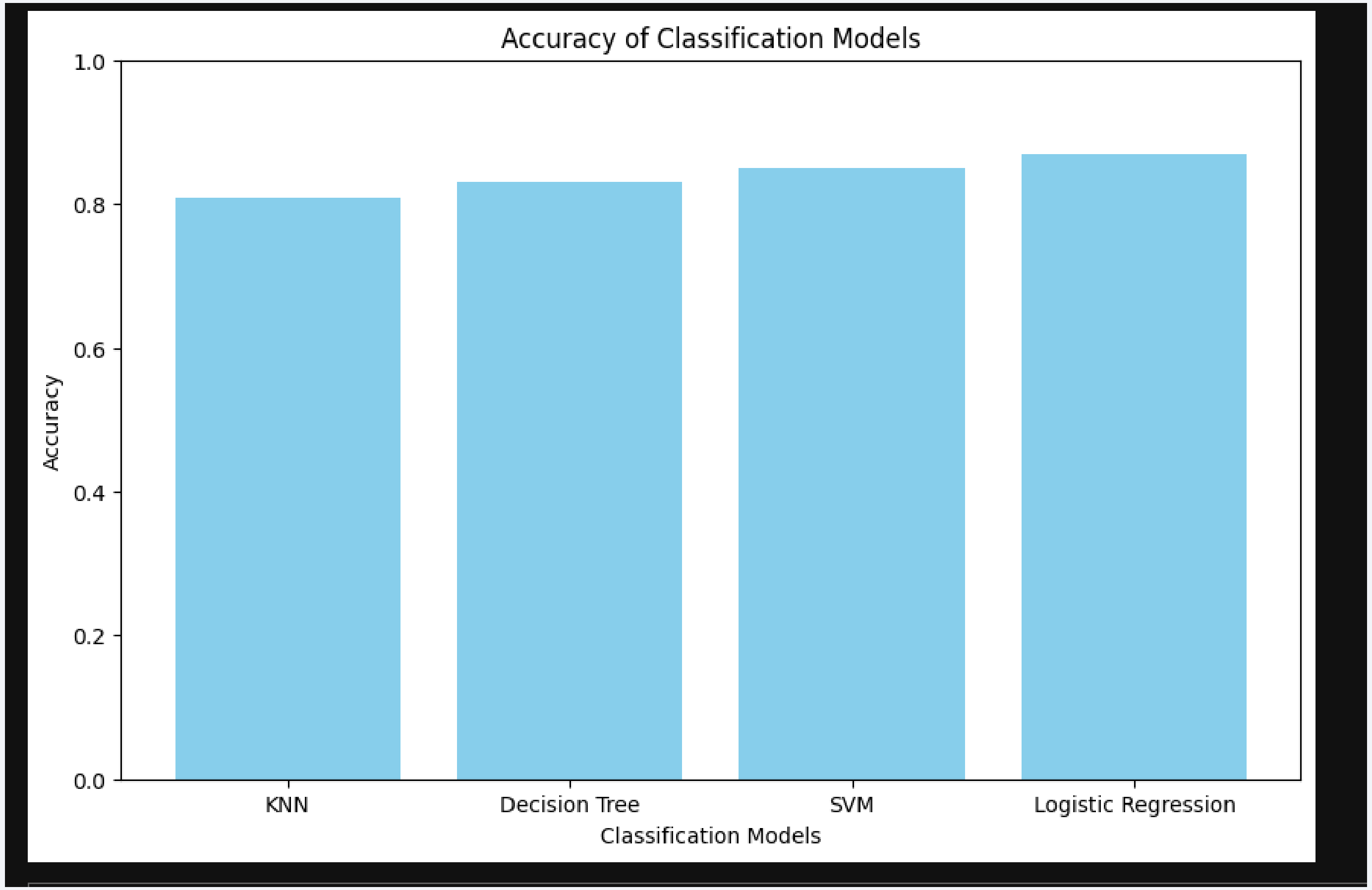
- Additionally, the scatter plot includes the Class feature, where 1 represents successful landings and 0 represents failures. It also incorporates the booster version category, represented by different colors, and the number of launches, reflected in the size of the data points.

- Remarkably, within the payload range of 0 to 6,000 kg, it's notable that there are two instances of failed landings despite the payload being listed as zero kilograms.

Section 5

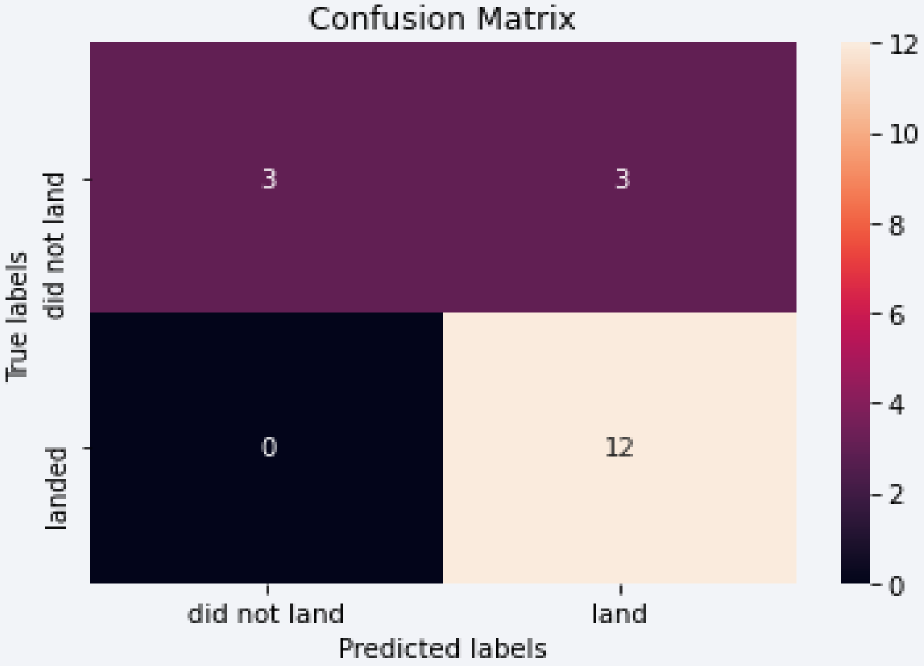
Predictive Analysis (Classification)

Classification Accuracy



- Logistic regression had the highest accuracy of 87%

Confusion Matrix



Conclusions

- Our objective is to develop a machine learning model for SpaceY to predict the successful landing of Stage 1 rockets, potentially saving approximately $100 million USD per successful landing. To accomplish this, we gathered data from a public SpaceX API and conducted web scraping of the SpaceX Wikipedia page. We then labeled the data and stored it in a DB2 SQL database. Additionally, we created a dashboard for visualizing the data.

- Our machine learning model achieved an accuracy of 87%. This means that Allon Mask of SpaceY can utilize this model to predict, with relatively high accuracy, whether a launch will result in a successful Stage 1 landing before the launch takes place. This prediction can assist in determining whether the launch should proceed or not, thus potentially saving significant costs.

- However, it's advisable to gather more data to further refine the machine learning model and improve its accuracy. Increasing the dataset size can enhance the model's performance and reliability, providing more robust predictions for SpaceY's decision-making process.

Thank you!