IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Arundhati Mondal
2 September

Outline

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

Executive Summary

- Summary of methodologies
  - API Data Collection
  - Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis
  - Interactive Dashboard with Shiny app
  - Predictive Analysis with Machine Learning

- Summary of all results
  - SpaceX launch success rates increase with payload mass.
  - Launches from KSC LC-39A have the highest success rates, making it an ideal location.
  - Decision Tree modeling proves most effective in predicting success.

Introduction

- Objective: Analyze SpaceX's launch data to inform SpaceY's competitive pricing strategy
- Methodology:
  - Data analysis of SpaceX's launch records
  - Identification of trends and correlations
  - Development of a predictive model
- Target Audience:
  - SpaceY leadership and stakeholders
  - Data science and engineering teams

Section 1

Methodology

Methodology

- Data collection methodology:
  - API
  - Web Scraping
- Data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection

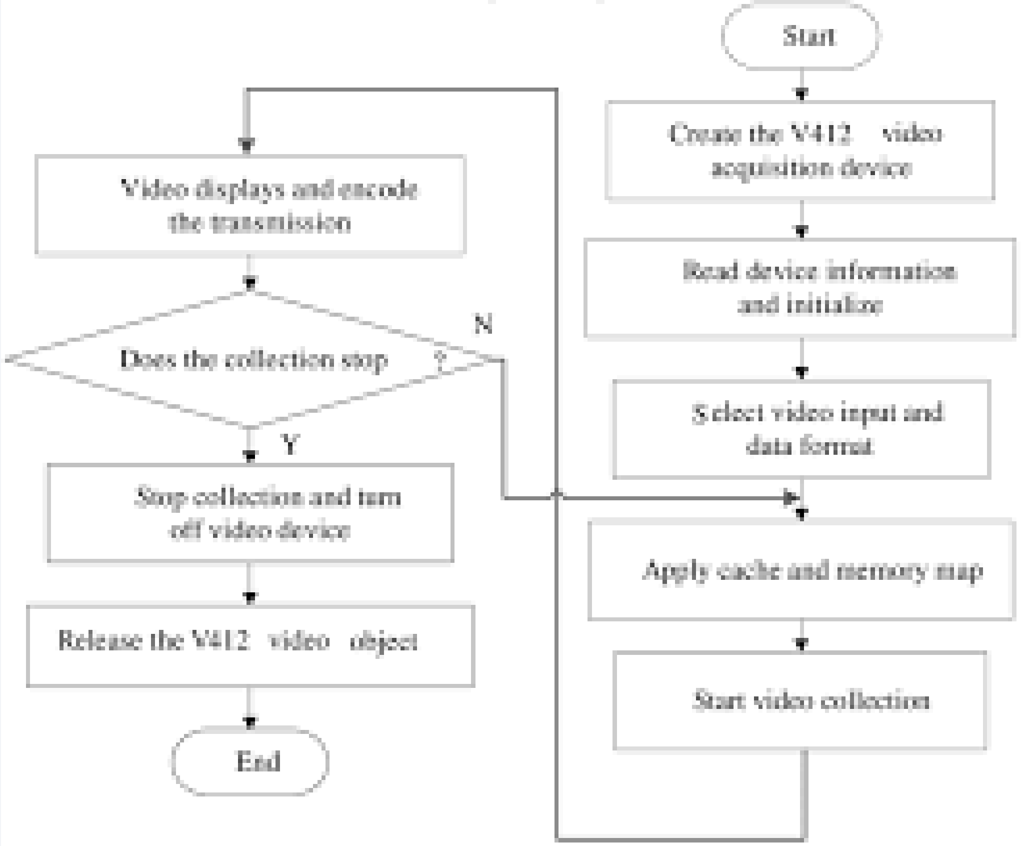
- SpaceX API Data:
  - Launch data for Falcon 9 rockets
  - Includes details such as:
  - Launch date and time
  - Mission name and description
  - Launch site and pad
  - Payload and customer information
  - Launch outcome (success/failure)
  - Rocket configuration and version

- Web Scraping from Wikipedia:
  - Additional data on Falcon 9 launches and rockets
  - Includes information such as:
  - Launch statistics and milestones
  - Rocket specifications (e.g., height, diameter, mass)
  - Engine and propulsion details
  - Payload capacity and fairing information
  - Launch history and timeline

Data Collection

- Using SpaceX API
  - Retrieve Falcon 9 launch data from SpaceX API via GET request
  - Decode API response into JSON format
  - Convert JSON data into a Pandas DataFrame
  - Extract relevant features for analysis
  - Handle missing data (replace/impute)

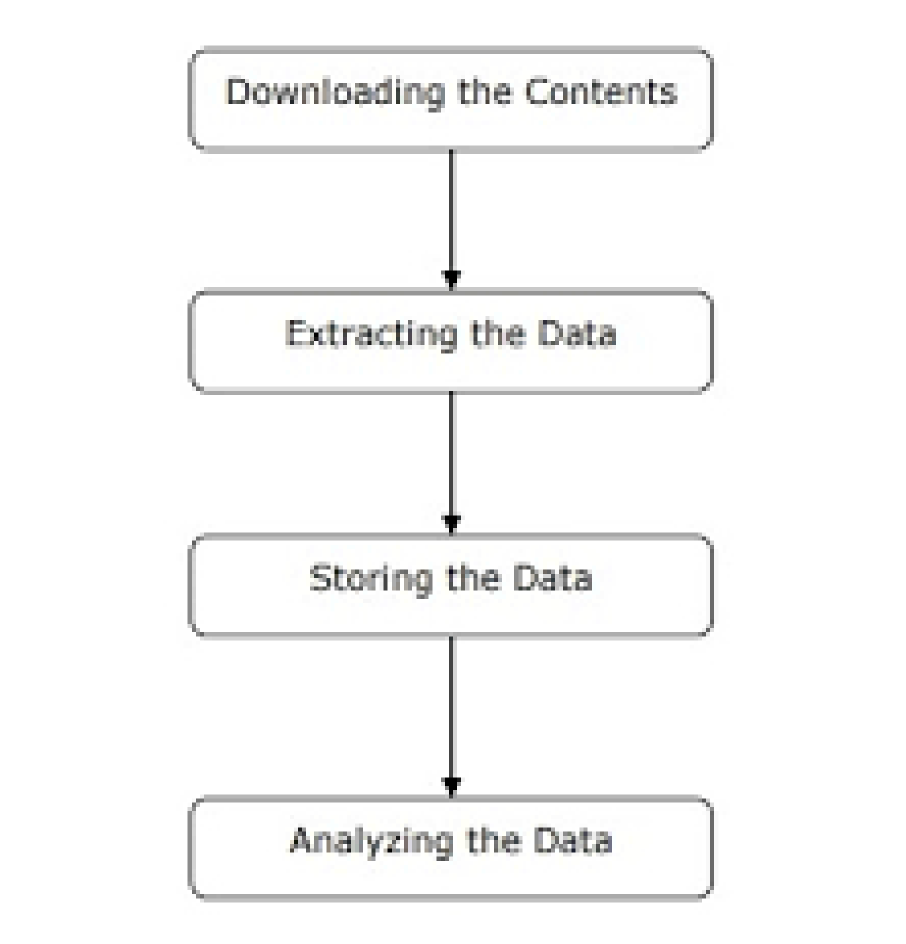
https://github.com/arundhati3700/Coursera/blob/main/Capstone/jupyter-labs-spacex-data-collection-api.ipynb


Start
Create the V412 video acquisition device
Read device information and initialize
Select video input and data format
Apply cache and memory map
Start video collection
Video displays and encode the transmission
Does the collection stop ?
N
Y
Stop collection and turn off video device
Release the V412 video object
End

Data Collection

- Using web scraping
  - Gather data through web scraping
  - Use BeautifulSoup to parse HTML content
  - Send GET request to retrieve data
  - Create DataFrame from HTML table
  - Extract table headers for column names

https://github.com/arundhati3700/Coursera/blob/main/Capstone/jupyter-labs-webscraping.ipynb


Downloading the Contents

Extracting the Data

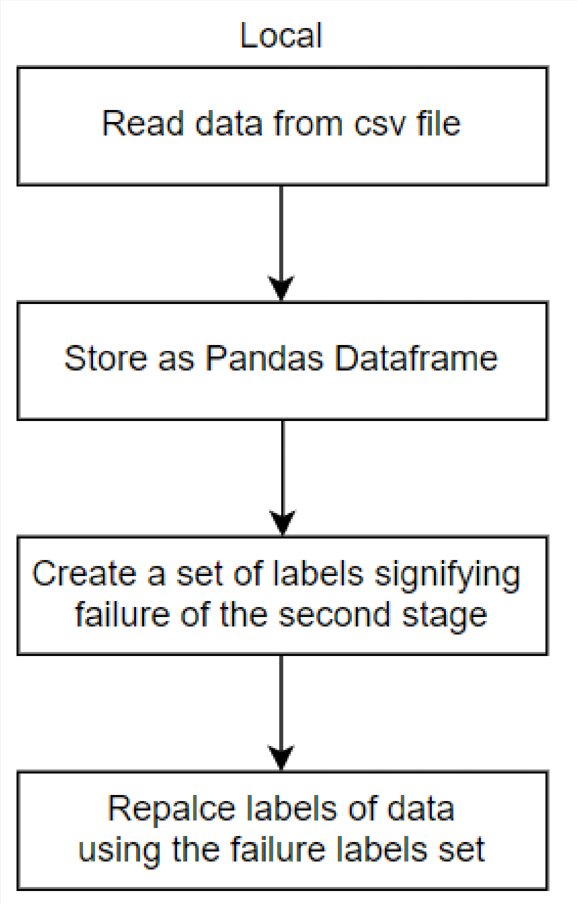
Storing the Data

Analyzing the Data

Data Wrangling

- Use collected data as input
- Main task: Standardize outcome descriptions
- Replace diverse outcome texts with:
  - Boolean values (True/False or 1/0)
  - Unified success/failure indicators

https://github.com/arundhati3700/Coursera/blob/main/Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb



Local

Read data from csv file

Store as Pandas Dataframe

Create a set of labels signifying failure of the second stage

Repalce labels of data using the failure labels set

EDA with Data Visualization

- Scatter Plots:
  - Flight Number vs. Launch Site (colored by outcome)
  - Payload Mass vs. Launch Site (colored by outcome)
  - Flight Number vs. Orbit Type (colored by outcome)
  - Payload Mass vs. Orbit Type (colored by outcome)
- Bar Chart:
  - Success rate of each orbit type (second stage)
  - Compares outcomes across orbit types (single numeric value per type)
- Line Plot:
  - Yearly trend of overall success rate
  - Shows temporal pattern of success and failure

https://github.com/arundhati3700/Coursera/blob/main/Capstone/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- date when the first succesful landing outcome in ground pad was acheived.
- names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- total number of successful and failure mission outcomes
- names of the booster_versions which have carried the maximum payload mass.
- records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- 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.

https://github.com/arundhati3700/Coursera/blob/main/Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Interactive map displays launch sites as circles and individual launches as markers
- Launches at the same site are clustered together
- Proximity analysis:
  - CCAFS SLC-40 launch site connected to:
    - Nearest coastline
    - Railway
    - Highway
    - City
- Insights:
  - CCAFS SLC-40 is a suitable launch site due to its:
    - Proximity to coastlines and transportation routes
    - Distance from populated cities

https://github.com/arundhati3700/Coursera/blob/main/Capstone/lab_jupyter_launch_site_location.ipynb

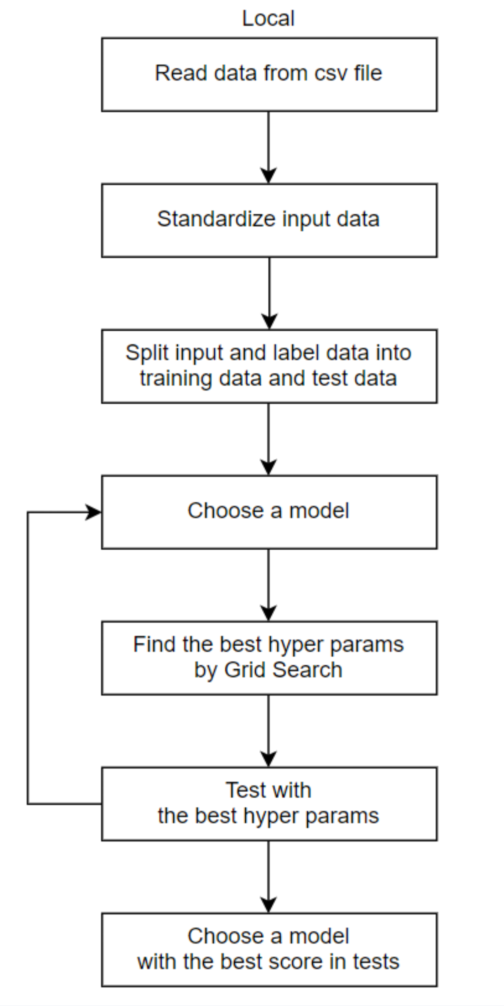
Build a Dashboard with Plotly Dash

- A pie chart and a scatter plot have been added on a dashboard
    - 5 different data sources can be selected interactively(4 sites and all sites)
- Pie chart: The landing success rate for each site and total success by site for all
- Scatter plot: The relationship between payload mass and landing success colored by Booster Version(Payload range can be changed)
- A pie chart can be used for comparing success rate of each site
- Trends in success based on Booster version and Payload mass can be seen in a scatter plot

https://github.com/arundhati3700/Coursera/blob/main/Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

- 4 machine learning methods compared: Logistic Regression, Support Vector Machine (SVM), Decision Tree, k-Nearest Neighbors (kNN)
- Model development process:
  - Data standardization
  - Data splitting
  - Model selection
  - Hyperparameter tuning with Grid Search
  - Testing
- Results:
  - Decision Tree performed best in this scenario
  - All methods had similar test scores
  - Decision Tree achieved highest scores in both training and testing
- https://github.com/arundhati3700/Coursera/blob/main/Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb


Local

Read data from csv file

Standardize input data

Split input and label data into training data and test data

Choose a model

Find the best hyper params by Grid Search

Test with the best hyper params

Choose a model with the best score in tests

Results

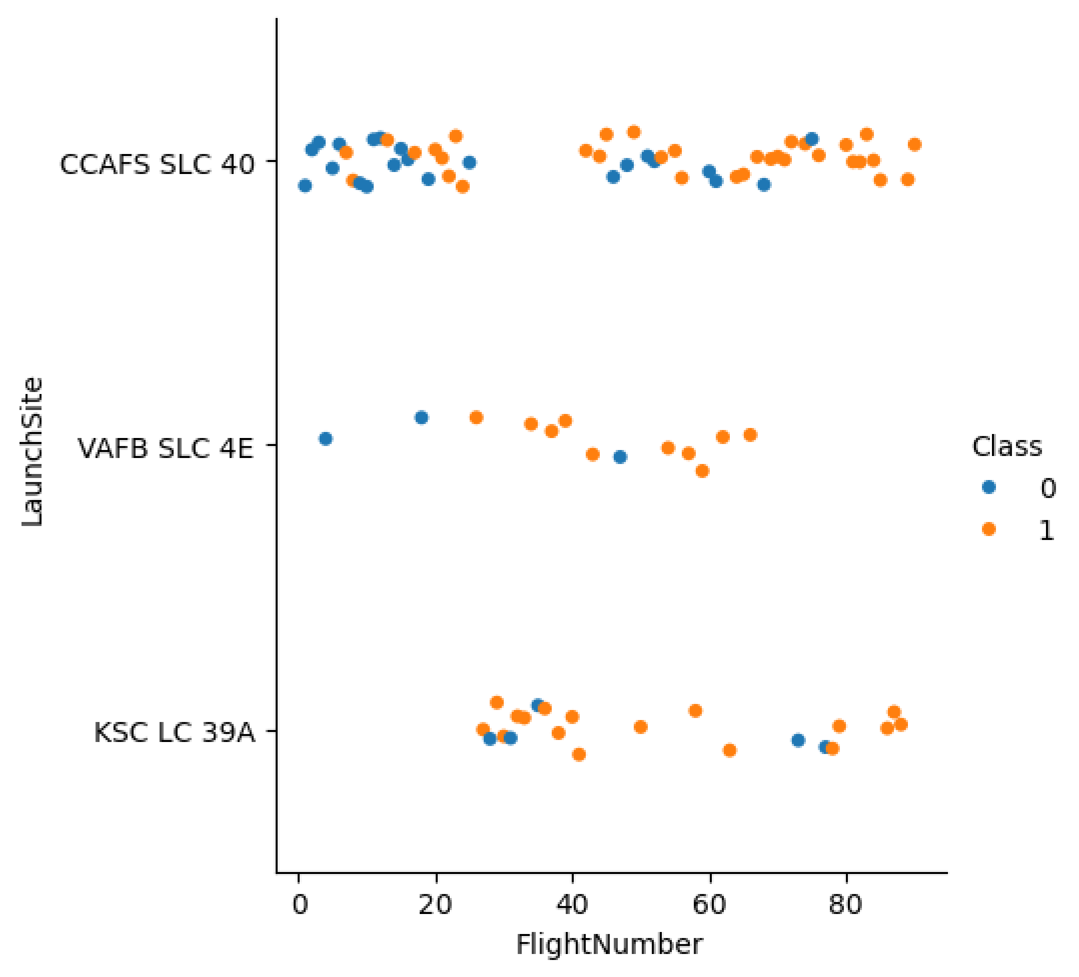
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

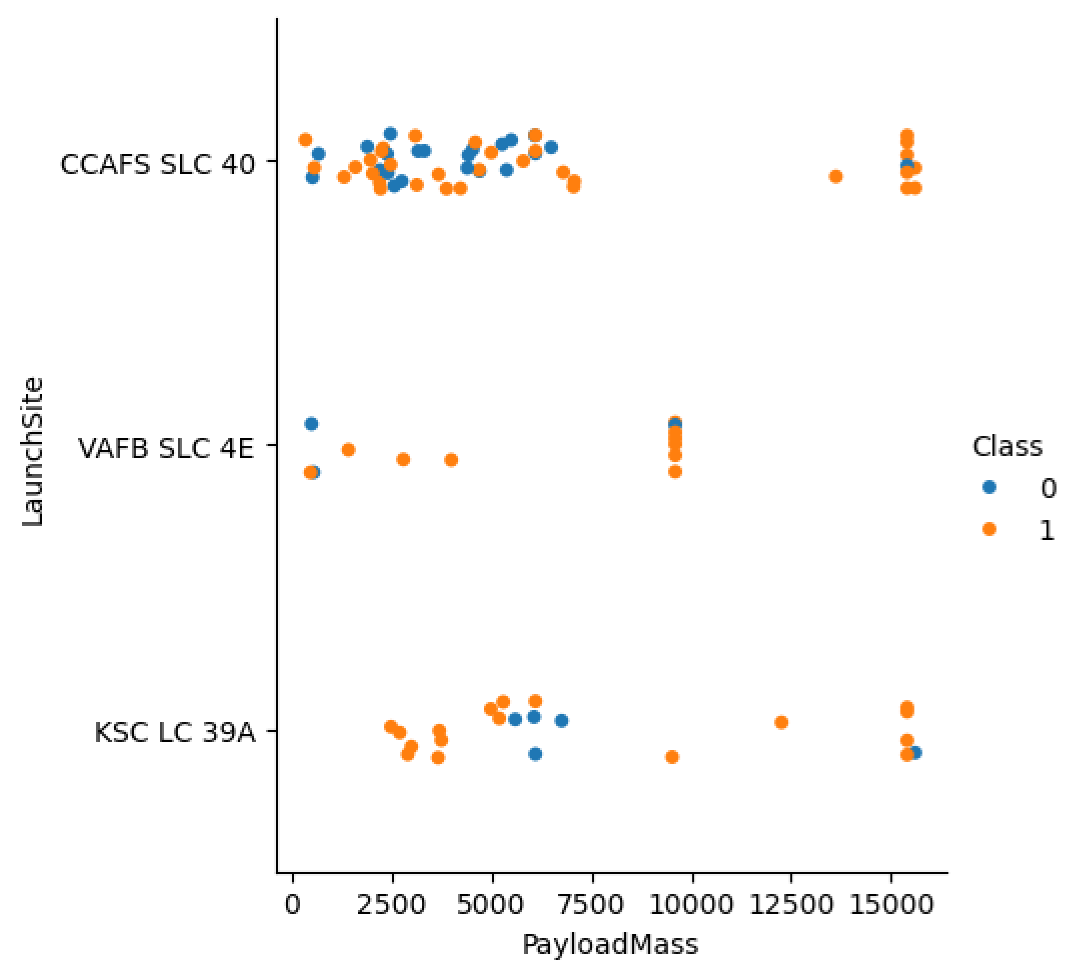
Flight Number vs. Launch Site

- The scatter plot of Flight Number vs. Launch Site reveals:
  - Launches at CCAFS SLC 40 exhibit a positive trend, where:
    - Higher Flight Numbers are associated with higher success rates
  - Launches at KSC LC 39A display a similar trend to CCAFS SLC 40, but with:
    - A higher overall success rate compared to CCAFS SLC 40
- This suggests that as the number of flights increases, the success rate tends to improve at both launch sites, with KSC LC 39A showing a slightly better performance.



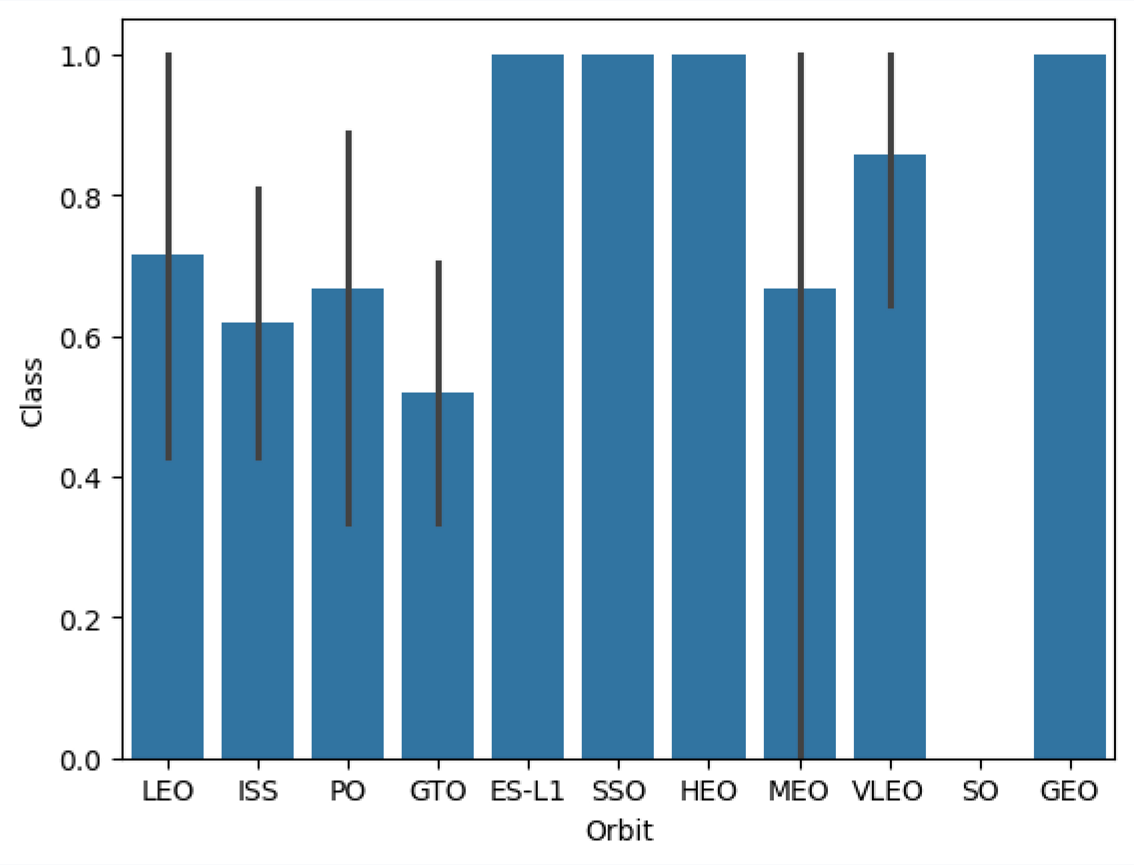
Payload vs. Launch Site

- The scatter plot of Payload Mass vs. Launch Site reveals:
  - High success rates are associated with Payload Masses above 10,000 kg
  - Launches at KSC LC 39A have a relatively high success rate, regardless of Payload Mass
- This suggests that heavier payloads tend to have higher success rates, and launches from KSC LC 39A are more likely to succeed, even with lighter payloads.



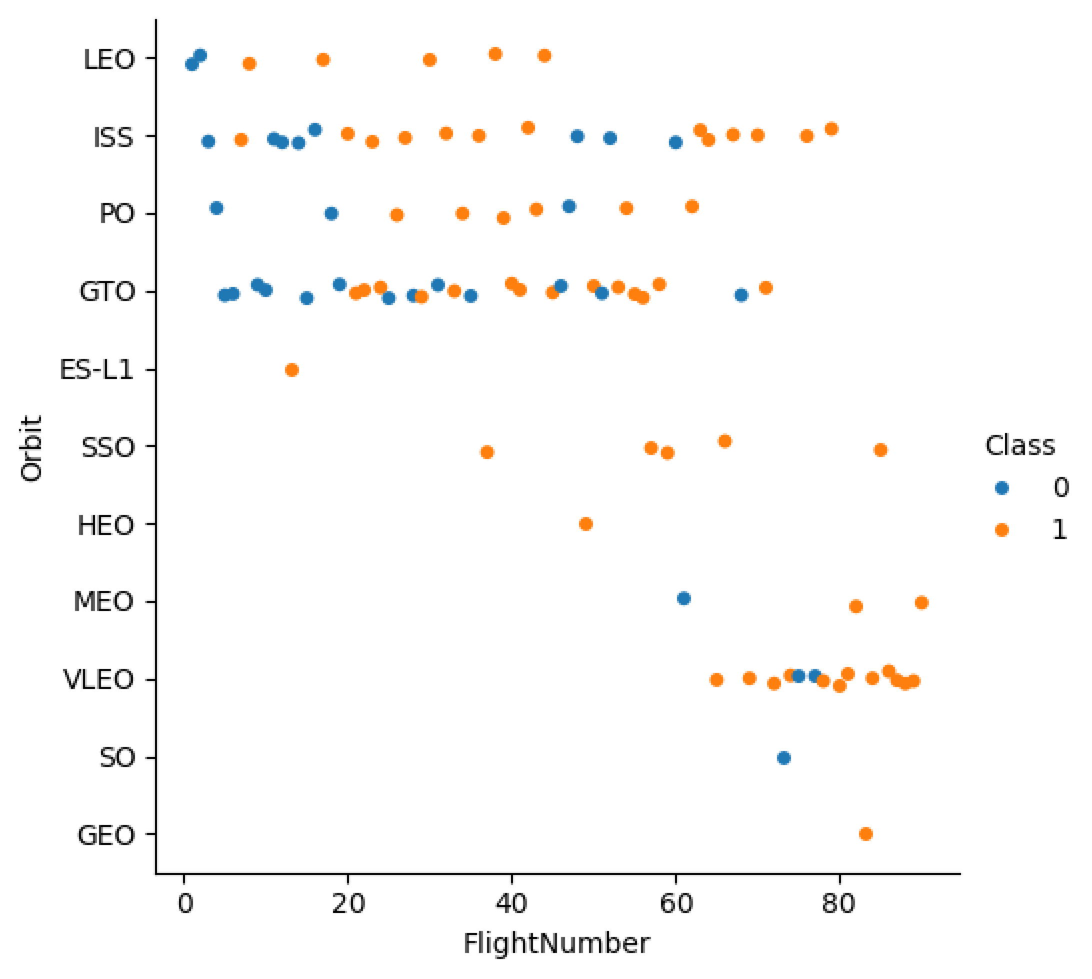
Success Rate vs. Orbit Type

- The bar chart shows the success rate of each orbit type:
  - ES-L1, SSO, HEO, and GEO orbit types have a 100% success rate
  - SO (Sun-Synchronous Orbit) cases have a 0% success rate, indicating failure every time
- This suggests that certain orbit types (ES-L1, SSO, HEO, and GEO) have a perfect track record, while SO orbit type launches have not been successful.



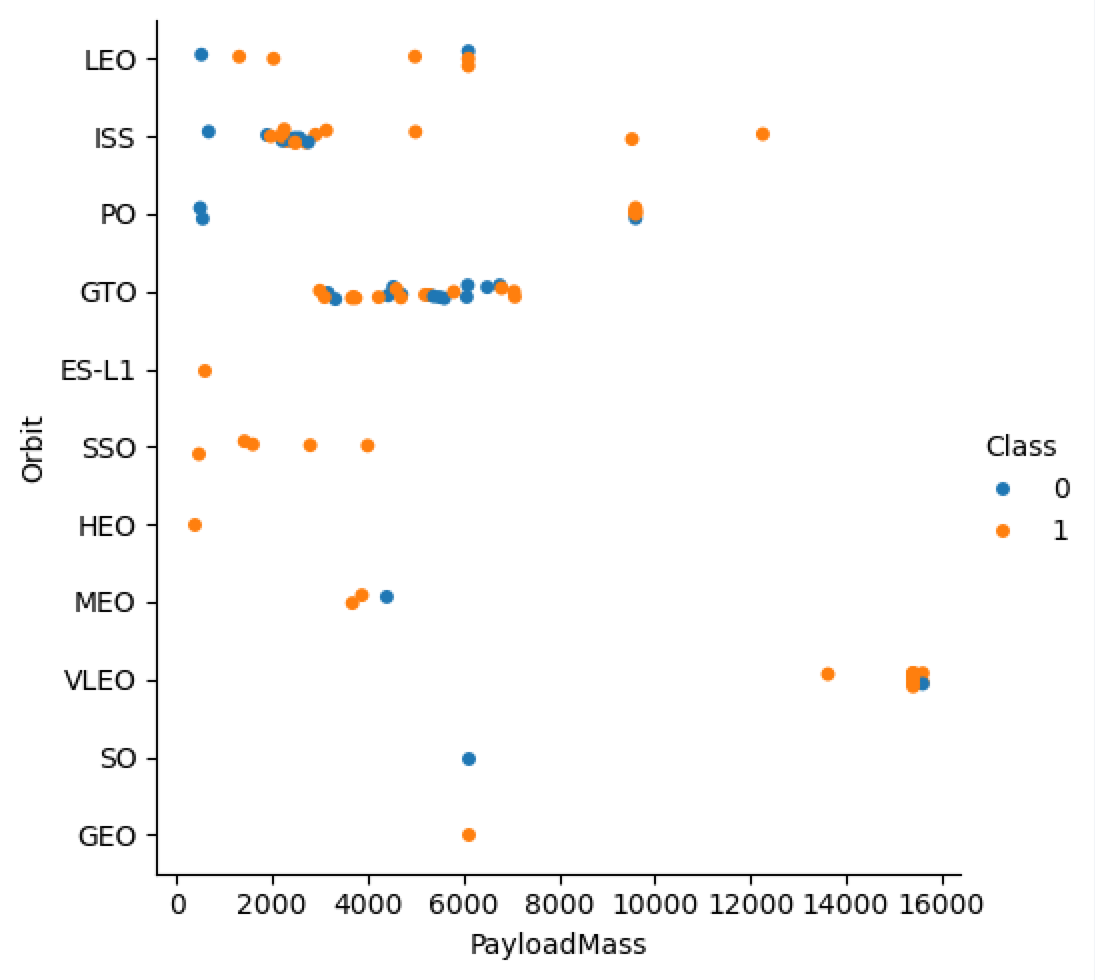
Flight Number vs. Orbit Type

- 100% success rate cases have a small number of attempts
- A trend emerges where later flights (higher Flight Numbers) tend to have higher success rates
- The number of flights increases, the success rate improves, and the orbit types with 100% success rates have been attempted fewer times.



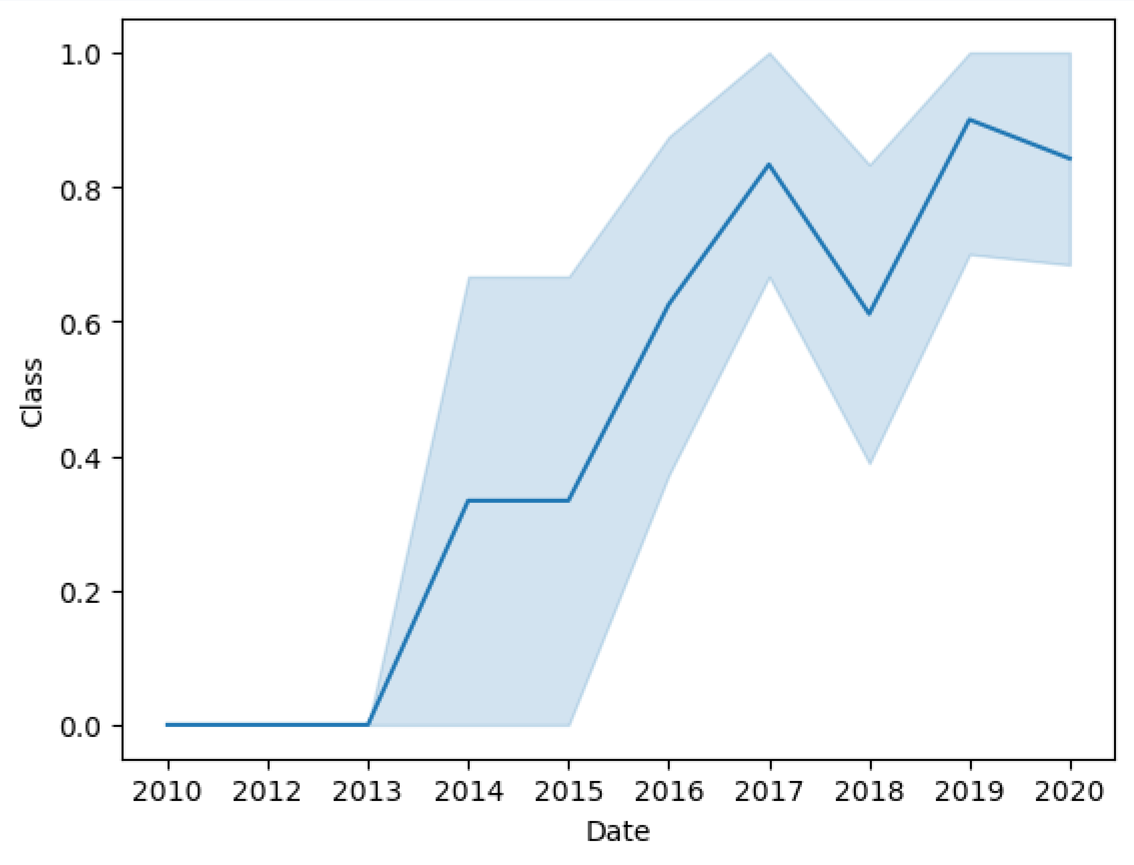
Payload vs. Orbit Type

- Higher Payload Mass is associated with a relatively higher success rate, except for GTO (Geostationary Transfer Orbit)
- However, the trend is not clear due to the limited number of launch attempts, making it difficult to draw a definitive conclusion
- While there may be a relationship between payload mass and success rate, the small sample size, especially for certain orbit types like GTO, makes it hard to confirm this trend.



Launch Success Yearly Trend

- A steady increase in success rate after 2013, indicating improvement over time
- However, the trend weakens after 2017, suggesting a slowdown in the rate of improvement
- This indicates that while there was a consistent improvement in success rates from 2013 to 2017, the progress has slowed down since then.



All Launch Site Names

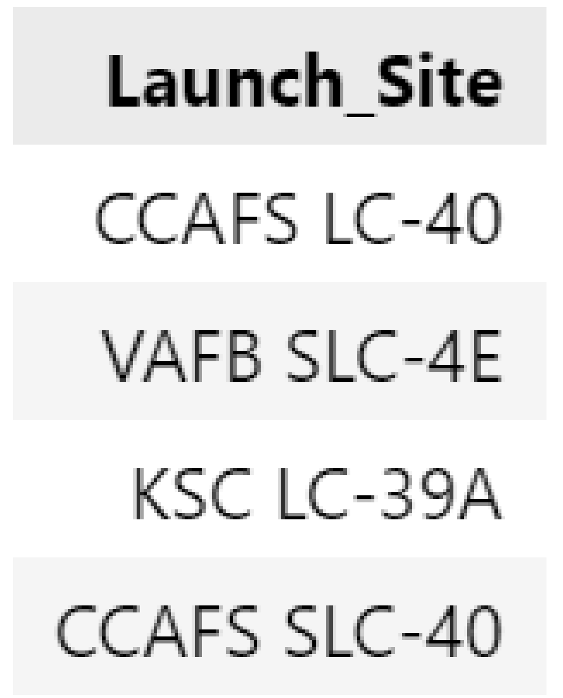
Find the names of the unique launch sites

Query:

%sql select distinct(Launch_Site) from SPACEXTBL

To obtain the unique launch sites names, distinct() command was used

This query eliminates duplicates and shows only the distinct launch site names, providing a concise list of all the different launch sites.



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

Launch Site Names Begin with 'CCA'

SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

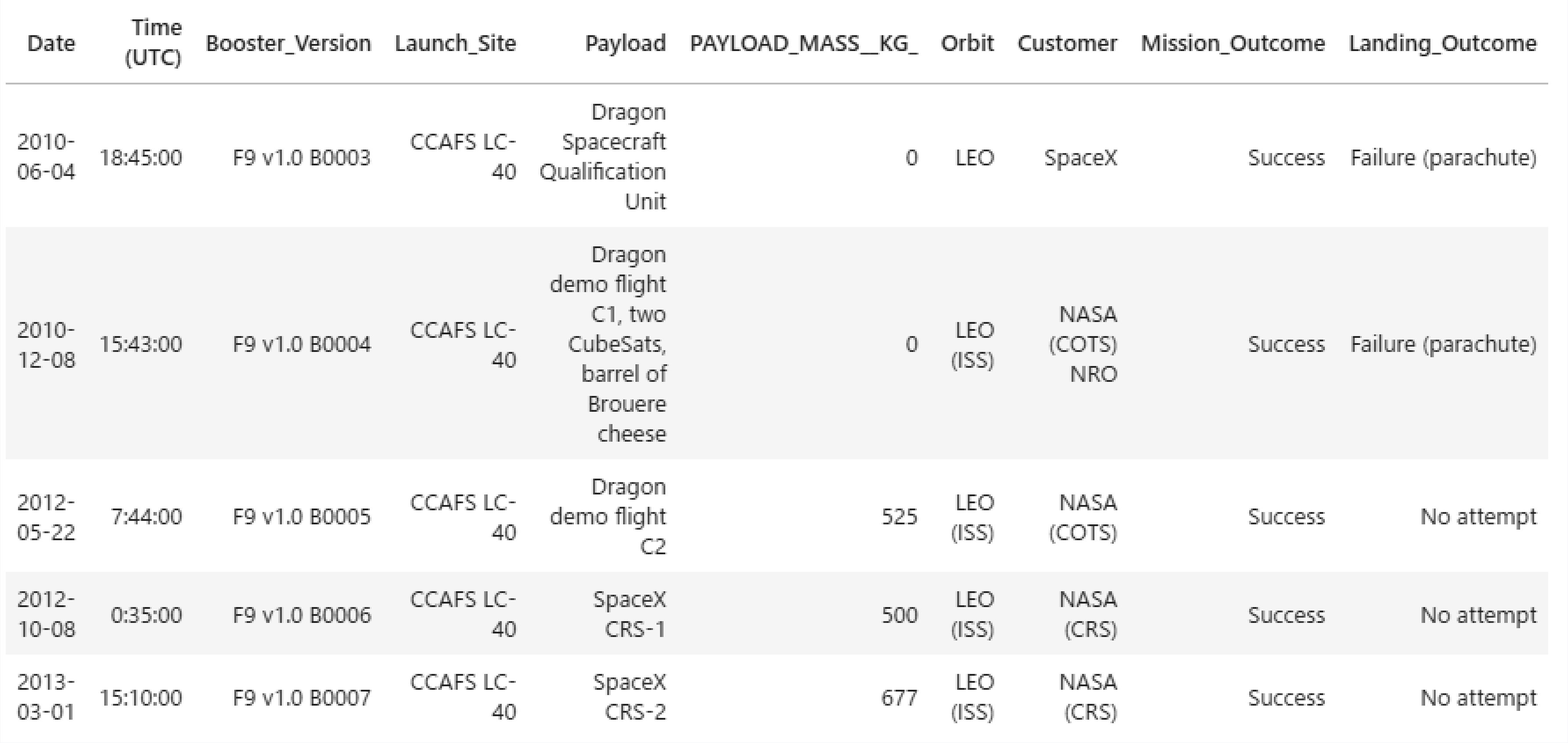
retrieves the first 5 records where the launch site name starts with the prefix 'CCA'.

Here's how the query works:

- SELECT * selects all columns from the table
- FROM SPACEXTBL specifies the table to query
- WHERE Launch_Site LIKE 'CCA%' filters records where the launch site name starts with 'CCA' (the '%' wildcard matches any characters after 'CCA')
- LIMIT 5 limits the output to the first 5 matching records

The result is a list of 5 records with launch site names starting with 'CCA', displayed on the right.

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

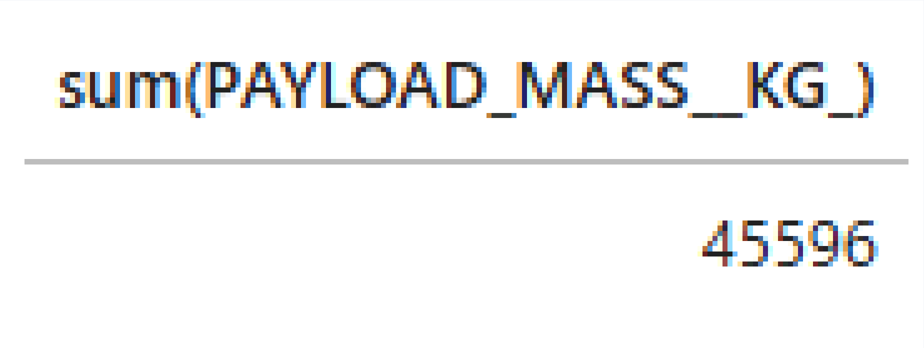
Total Payload Mass

SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'

Calculates the total payload mass carried by boosters for NASA (CRS) customers.

Here's how the query works:

- SELECT SUM(PAYLOAD_MASS__KG_) calculates the sum of the payload mass values
- FROM SPACEXTBL specifies the table to query
- WHERE Customer = 'NASA (CRS)' filters records where the customer is NASA (CRS)



sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE '%F9 v1.1%'

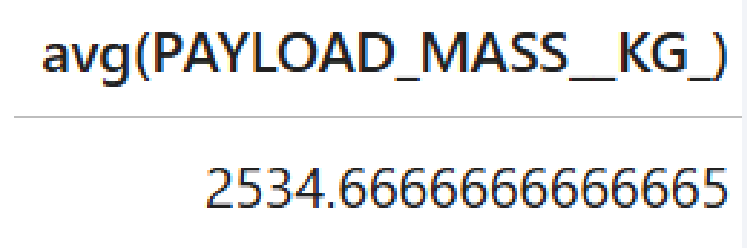
calculates the average payload mass carried by the F9 v1.1 booster version.

Here's how the query works:

- SELECT AVG(PAYLOAD_MASS__KG_) calculates the average of the payload mass values

- FROM SPACEXTBL specifies the table to query

- WHERE Booster_Version LIKE '%F9 v1.1%' filters records where the booster version is F9 v1.1 (the '%' wildcard matches any characters before and after 'F9 v1.1')



avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'

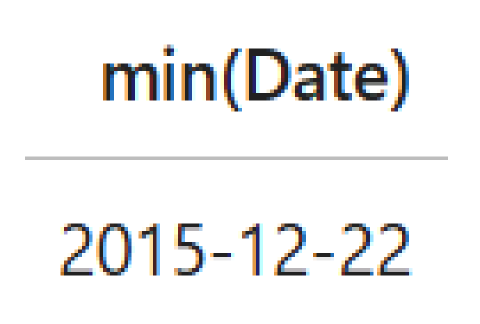
retrieves the date of the first successful landing outcome on a ground pad.

Here's how the query works:

- SELECT MIN(Date) returns the minimum date value (i.e., the earliest date)

- FROM SPACEXTBL specifies the table to query

- WHERE Landing_Outcome = 'Success (ground pad)' filters records where the landing outcome was a success on a ground pad



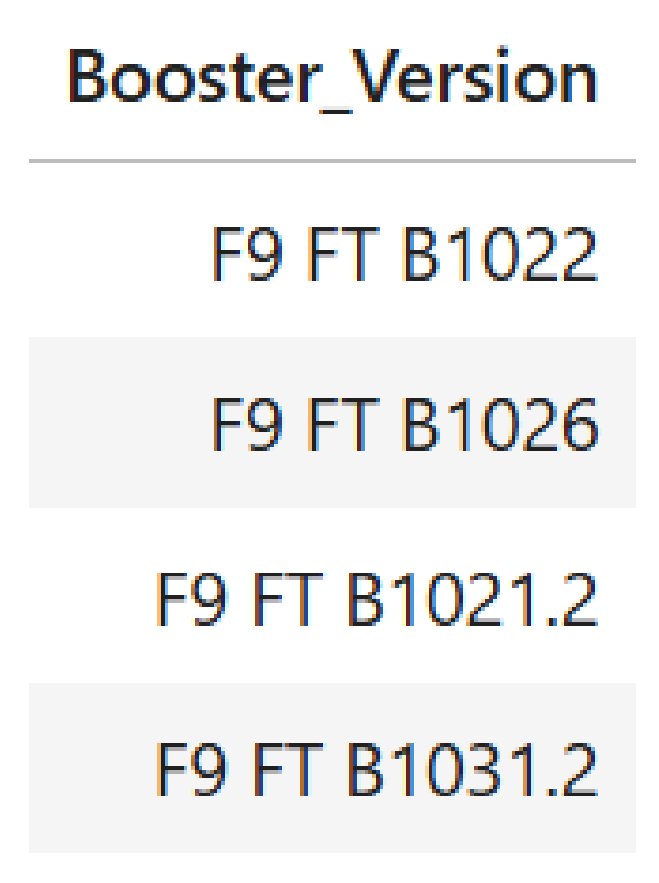
min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version FROM SPACEXTBL WHERE
   Landing_Outcome = 'Success (drone ship)' AND
   PAYLOAD_MASS__KG_ > 4000 AND
   PAYLOAD_MASS__KG_ < 6000
```

retrieves the names of boosters that meet the following conditions:

1. Successfully landed on a drone ship (Landing_Outcome = 'Success (drone ship)')
2. Had a payload mass greater than 4000 kg (PAYLOAD_MASS__KG_ > 4000)
3. Had a payload mass less than 6000 kg (PAYLOAD_MASS__KG_ < 6000)



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

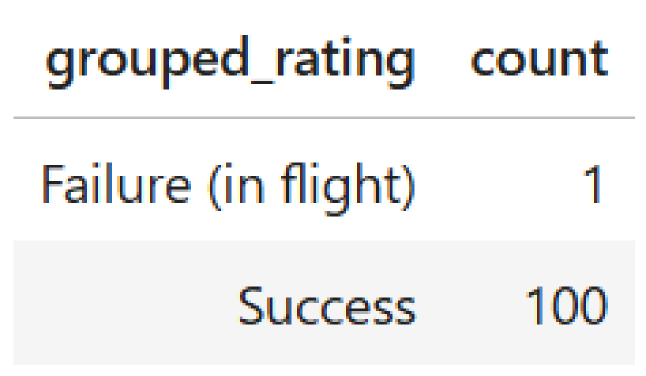
Total Number of Successful and Failure Mission Outcomes

The SQL query:

```
SELECT CASE WHEN Mission_Outcome LIKE '%Success%'
   THEN 'Success' ELSE 'Failure (in flight)' END AS
   grouped_rating, COUNT(*) AS count FROM SPACEXTBL
   GROUP BY grouped_rating;
```

retrieves the total number of successful and failure mission outcomes that meet the following conditions:

1. Mission outcome contains the word 'Success' (Mission_Outcome LIKE '%Success%')
2. Mission outcome does not contain the word 'Success' (ELSE 'Failure (in flight)')



grouped_rating	count
Failure (in flight)	1
Success	100

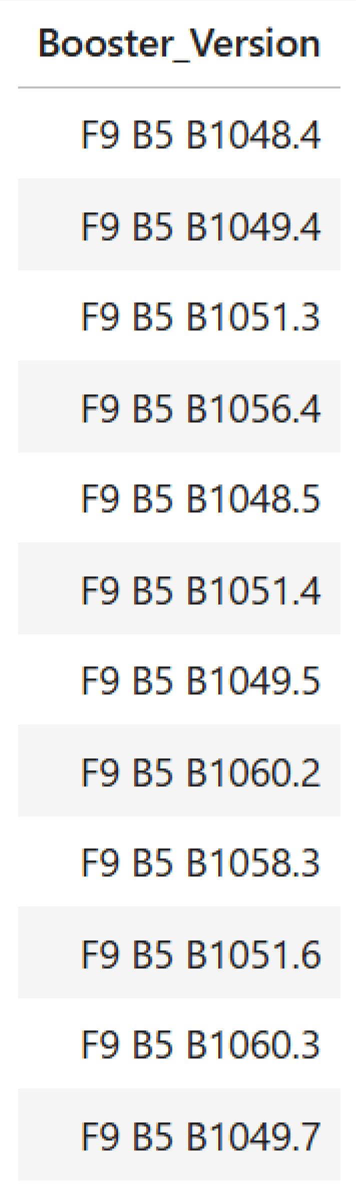
Boosters Carried Maximum Payload

The SQL query:

SELECT DISTINCT(Booster_Version) FROM SPACEXTBL WHERE PAYLOAD_MASS_KG = (SELECT MAX(PAYLOAD_MASS_KG) FROM SPACEXTBL)

retrieves the names of boosters that have carried the maximum payload mass, meeting the following condition:

1. Payload mass is equal to the maximum payload mass recorded in the SPACEXTBL database (PAYLOAD_MASS_KG = (SELECT MAX(PAYLOAD_MASS_KG) FROM SPACEXTBL))



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

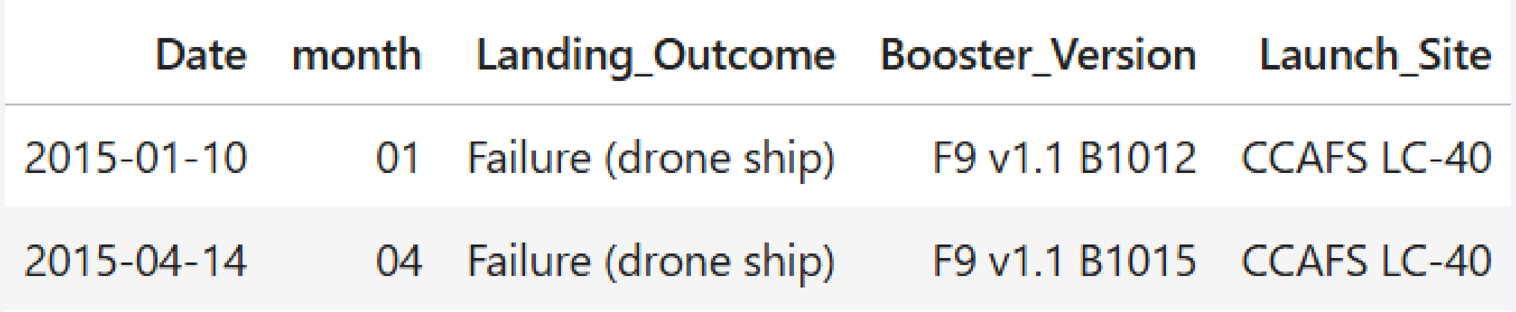
2015 Launch Records

The SQL query:

```
SELECT Date, SUBSTR(Date, 6,2) AS month, Landing_Outcome,
   Booster_Version, Launch_Site FROM SPACEXTBL WHERE SUBSTR(Date, 0,5) =
   '2015' AND Landing_Outcome = 'Failure (drone ship)'
```

retrieves the following information for failed drone ship landings in 2015:

Date of launch, Month of launch (extracted from the Date), Landing outcome (specifically, 'Failure (drone ship)'), Booster version, Launch site name



Date	month	Landing_Outcome	Booster_Version	Launch_Site
2015-01-10	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

The SQL query:

```
SELECT Landing_Outcome, COUNT(*) AS count_ FROM SPACEXTBL WHERE Date >
'2010-06-04' AND Date < '2017-03-20' GROUP BY Landing_Outcome ORDER BY
count_ DESC
```

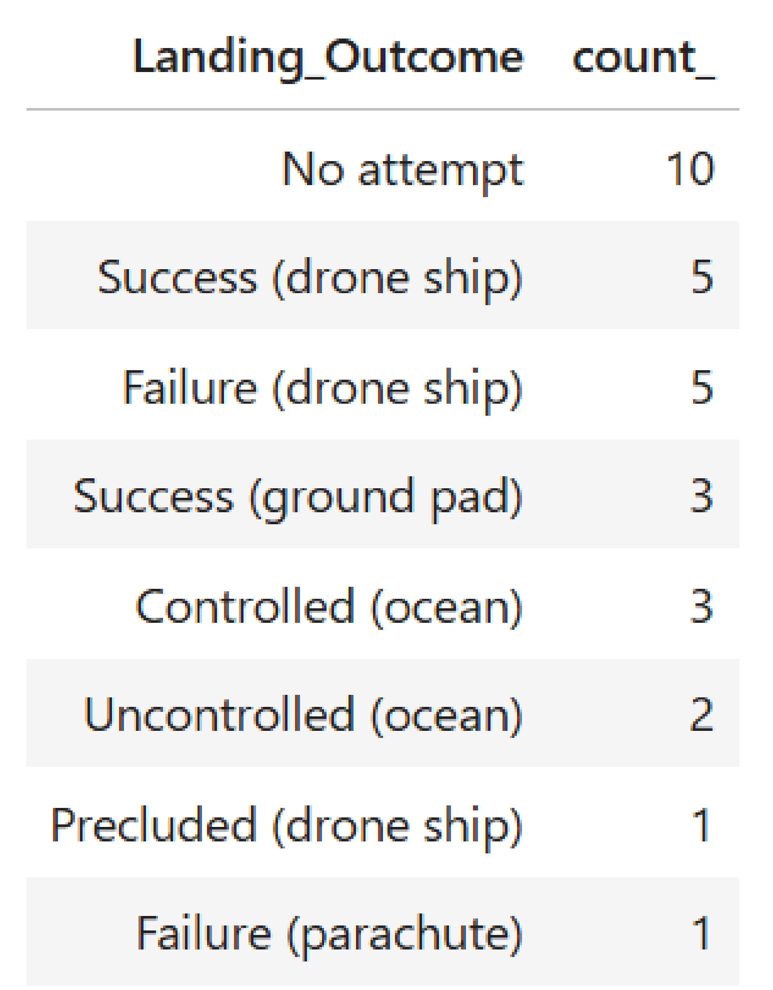
retrieves the count of landing outcomes between June 4, 2010, and March 20, 2017, grouped by outcome and sorted in descending order. The result shows that:

- The most frequent landing outcome is attempts to land on a drone ship (excluding 'No attempt' cases)

- The count of each landing outcome is displayed in descending order, with the highest count first

This indicates that drone ship landings were the most common type of landing attempt during this time period.

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



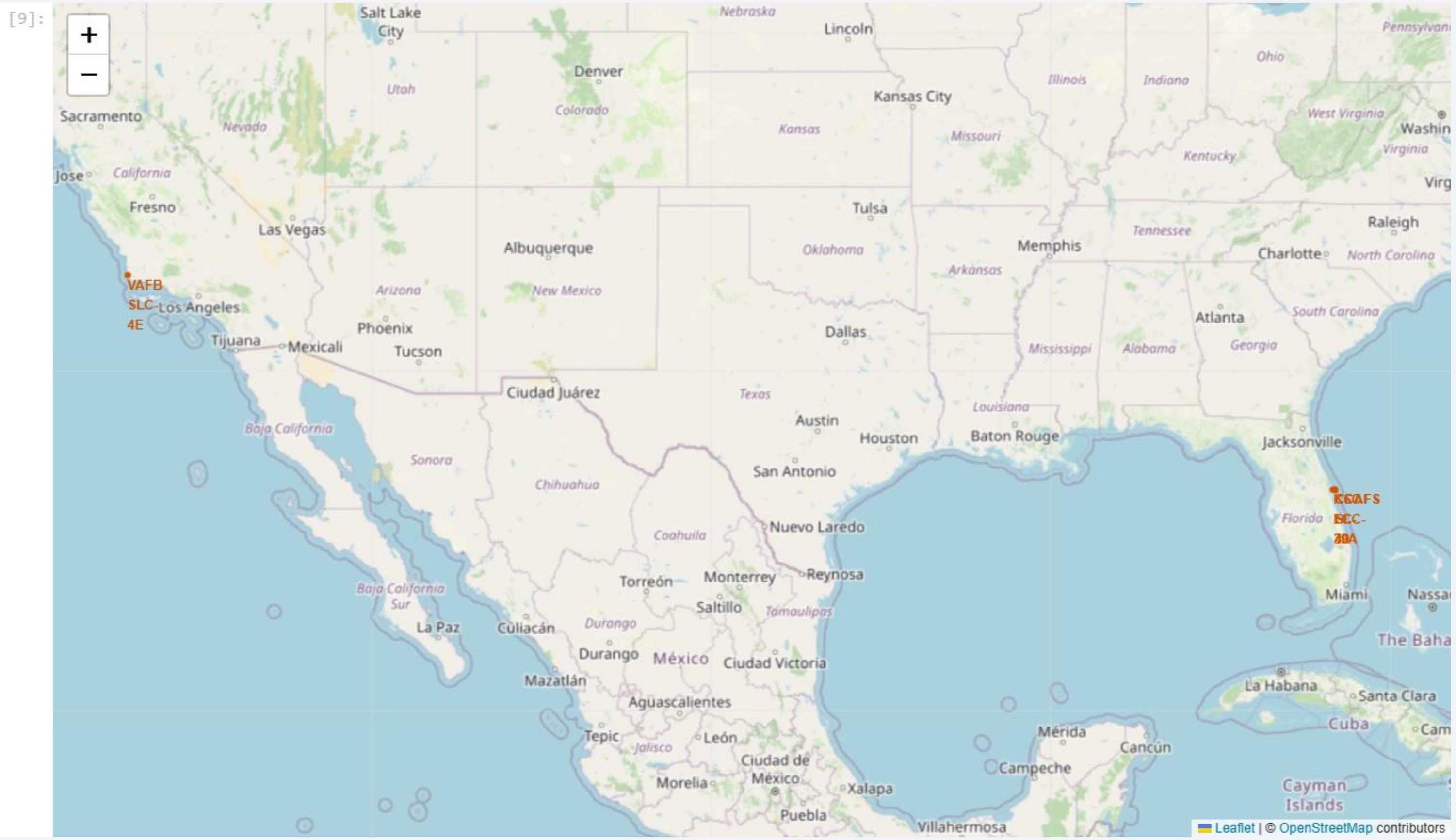
Landing_Outcome	count_
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites in the United States

The figure on the right displays the locations of all launch sites in the United States, marked with red markers. Notably, all of these launch sites are situated in coastal areas or peninsulas, indicating a strategic preference for launching spacecraft from these types of locations.



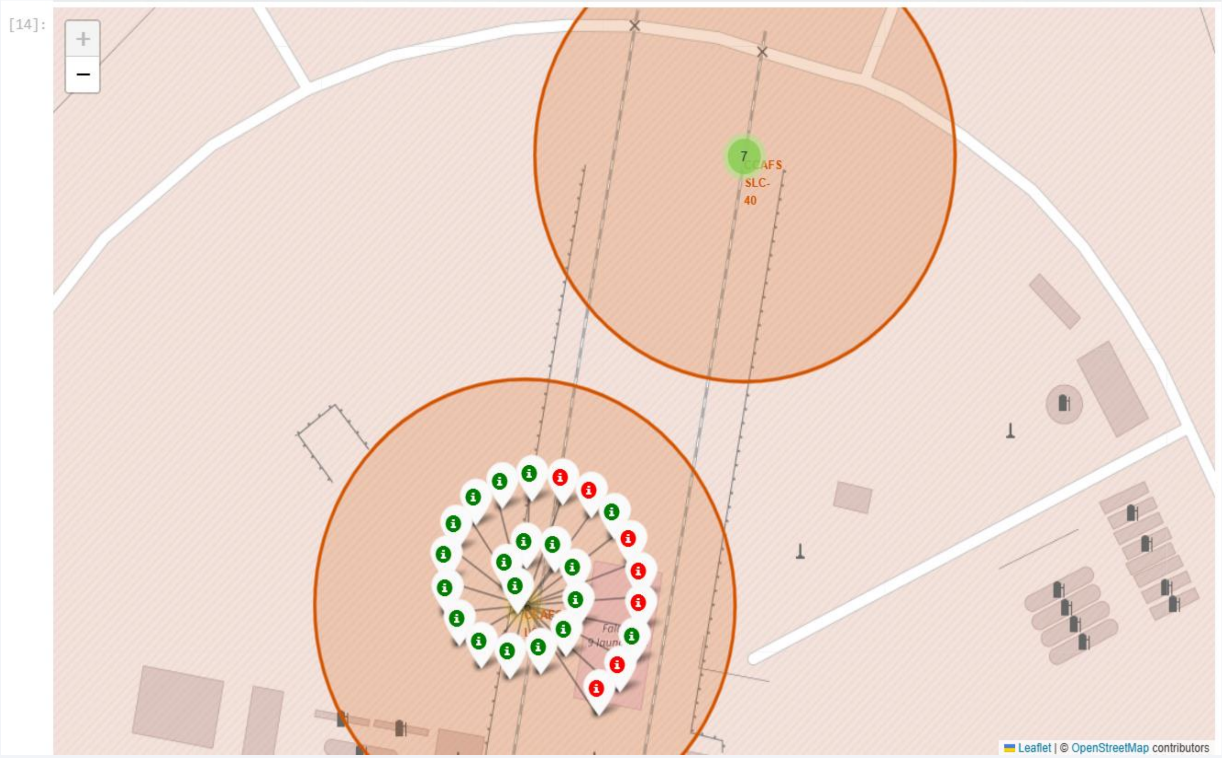
Color-labeled Launch Outcomes on the Map

The figure on the right illustrates the launch outcomes at CCFAS LC-40 using a spiral dot sequence, where:

- Green markers indicate successful landings

- Red markers indicate failed landings

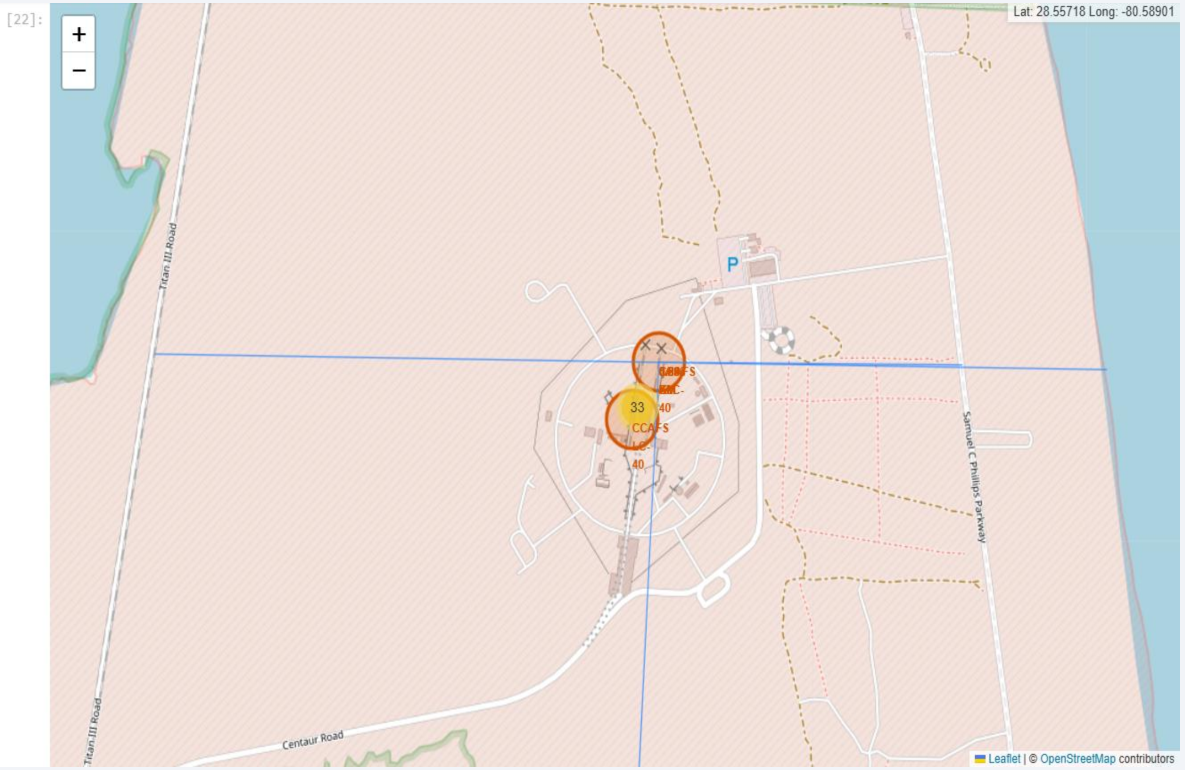
The pattern shows that launches at CCFAS LC-40 have a high probability of successfully landing, as evidenced by the predominance of green markers. This suggests a strong track record of successful missions at this launch site.



Distances between a Launch Site and Important Proximities

The figure on the right displays the distances from the launch site to nearby features using blue lines, which indicate that:

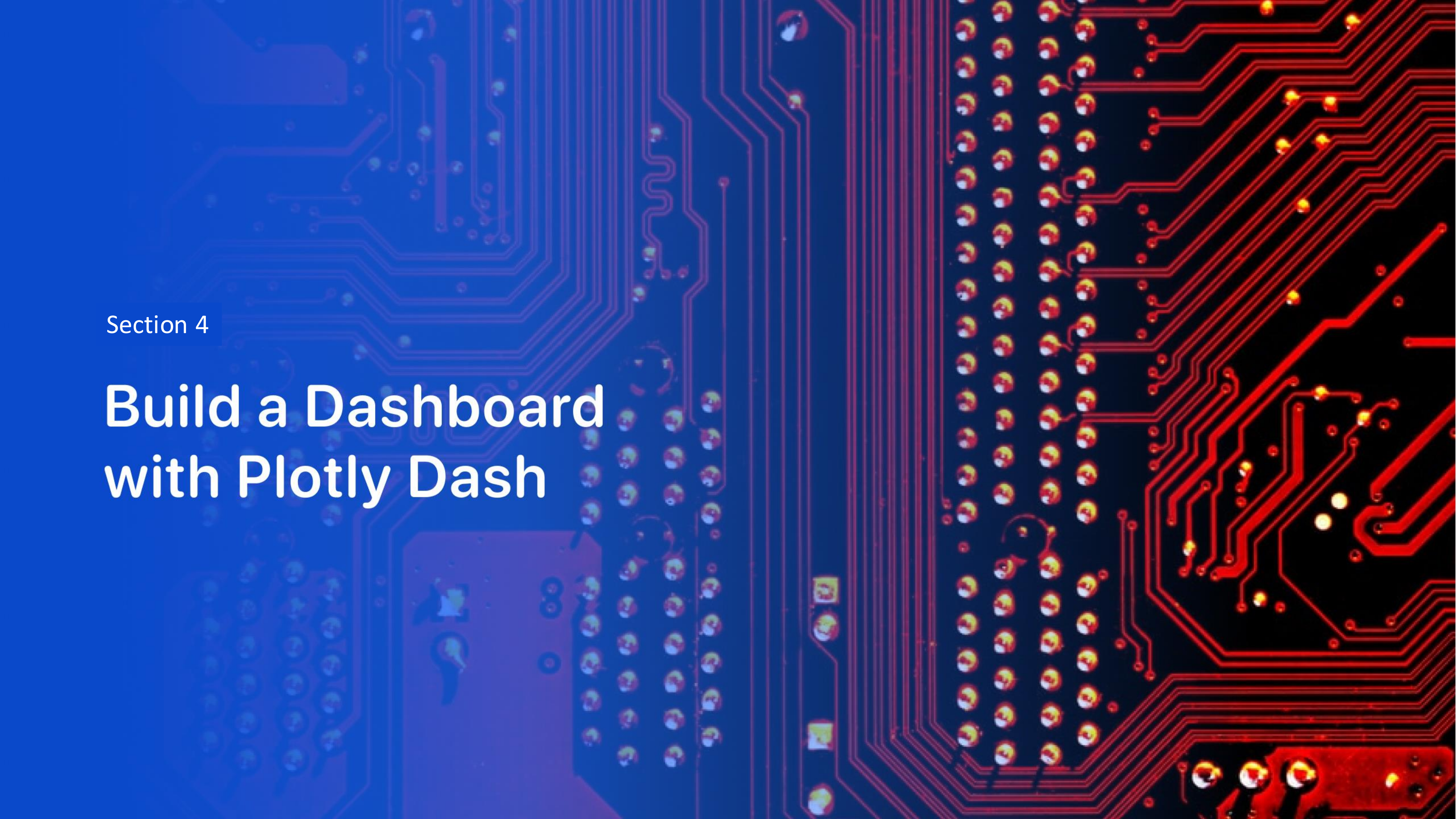
- The launch site is close to a highway (left side, closer proximity)
- The launch site is also near a coastline (left side, further proximity)
- The launch site has a railway nearby (right side)


[22]:
+
−
Lat: 28.55718 Long: -80.58901
Titan III Road
Titan III Road
P
33
CCAFS LC-40
CCAFS LC-40
Samuel C Phillips Parkway
Centaur Road
Leaflet | © OpenStreetMap contributors

The short lengths of the blue lines suggest that these features are relatively close to the launch site, indicating a strategically located launch site with convenient access to transportation infrastructure.

Section 4

Build a Dashboard with Plotly Dash

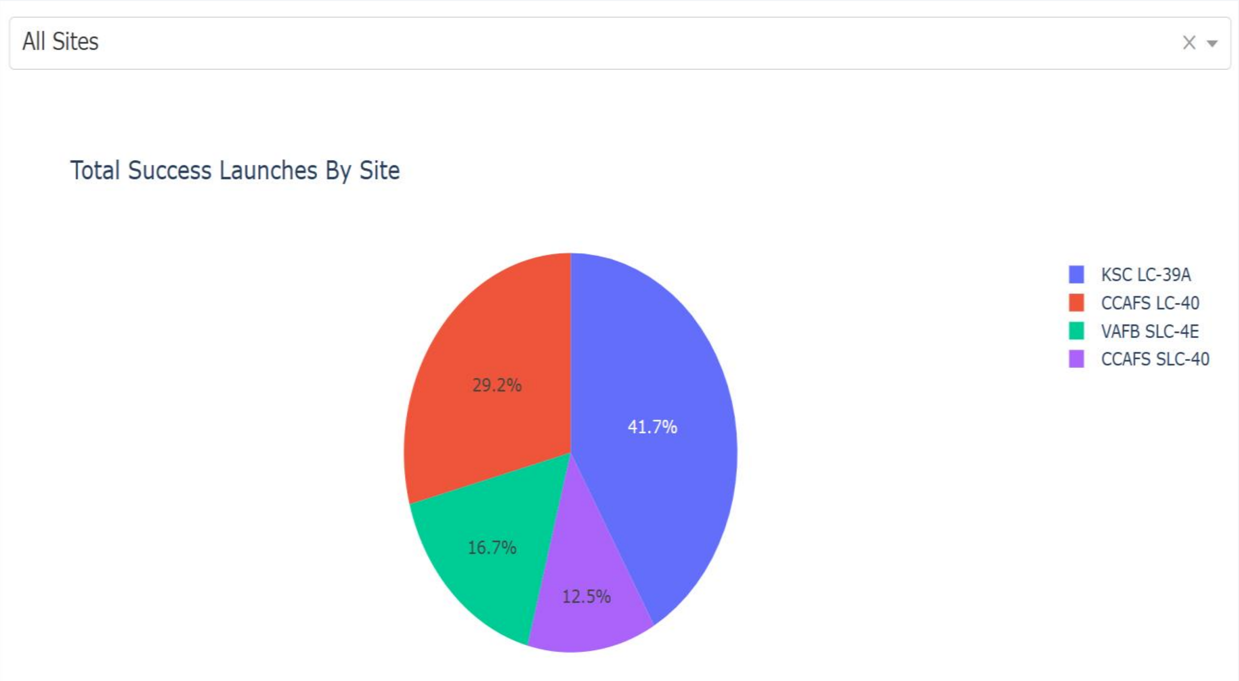


Proportion of Each Site in the Total Success Launches

The pie chart illustrates the distribution of successful launches among various sites, revealing that:

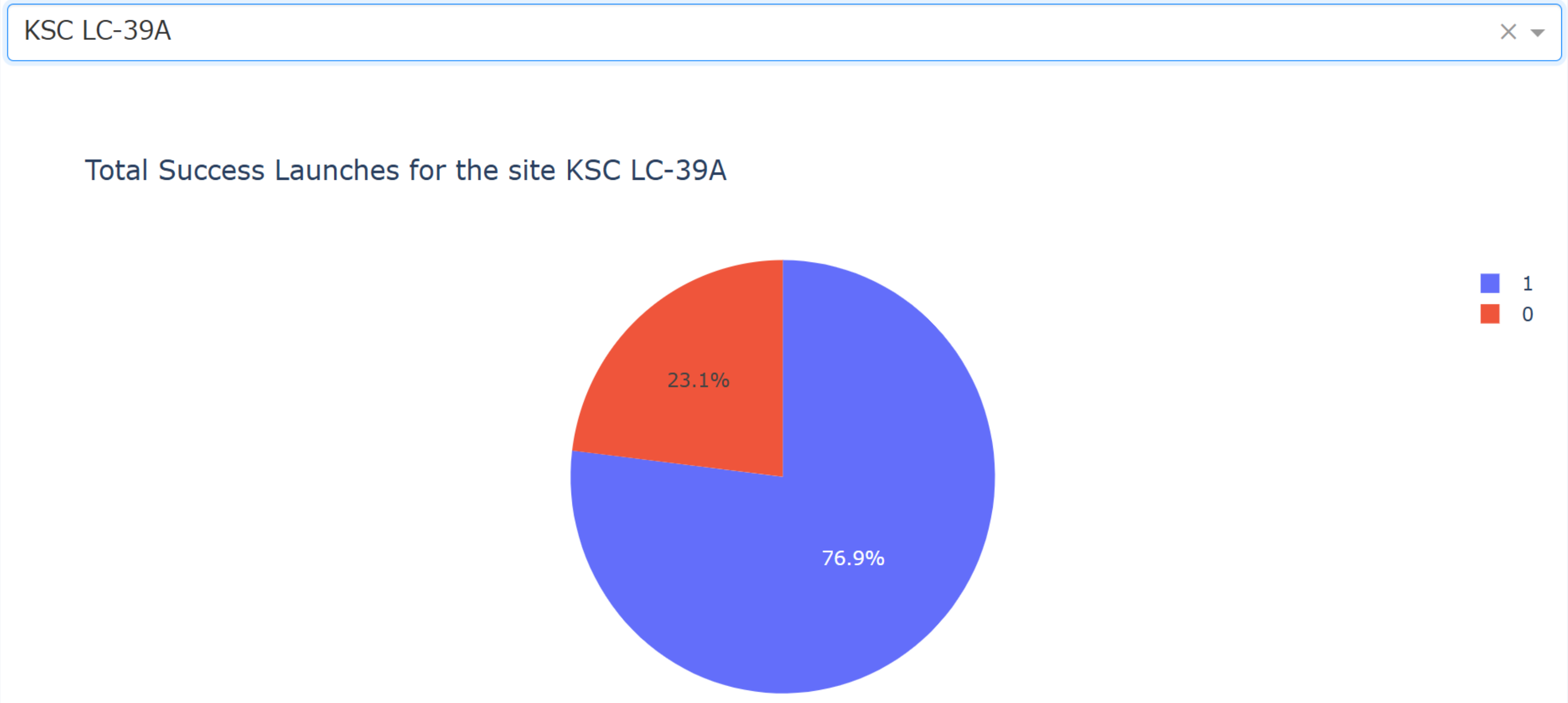
- KSC LC-39A has the largest share of successful launches, making up the biggest proportion of the total
- Other sites have smaller percentages of successful launches, as indicated by the smaller slices of the pie chart

This suggests that KSC LC-39A is the most successful launch site, accounting for the majority of successful launches.



Launch Success Rate at KSC LC-39A

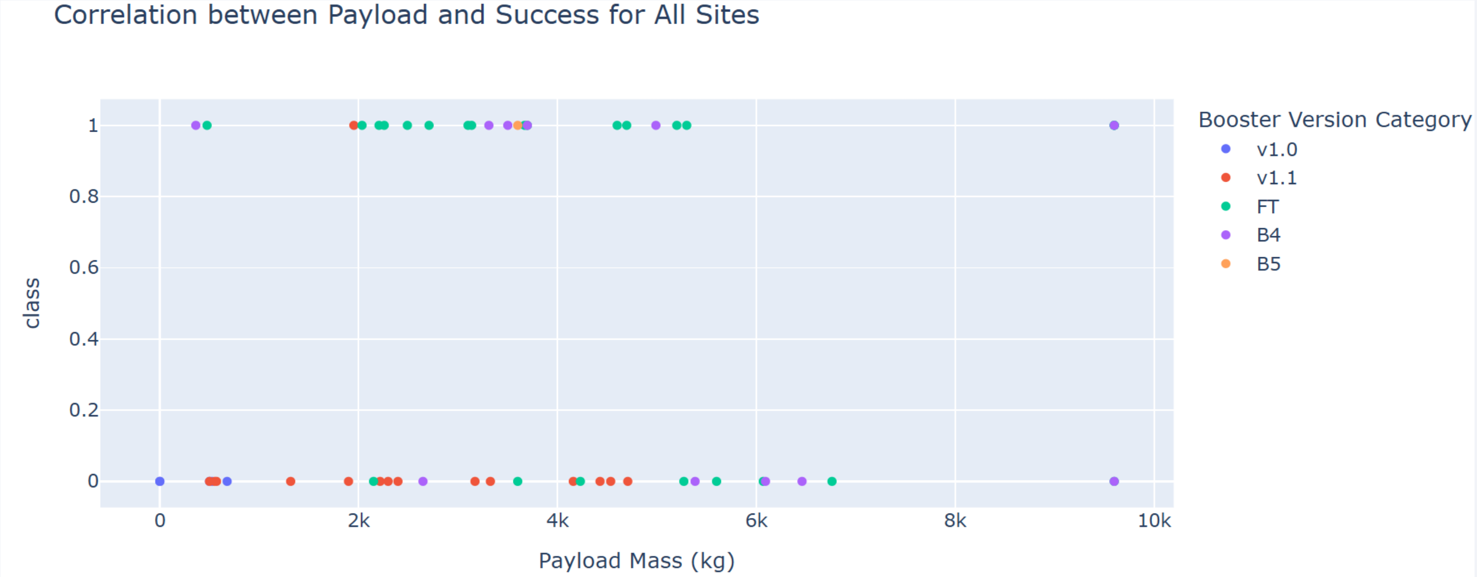
The pie chart displays the success rate at KSC LC-39A, which boasts the highest launch success ratio, showcasing its impressive performance.


KSC LC-39A

Total Success Launches for the site KSC LC-39A

Launch Outcomes Labeled with Booster Version

A scatter plot at the bottom illustrates the relationship between payload mass and launch outcome for all sites, with a range slider to select different payloads. The plot reveals a high success rate when using the FT booster, particularly when the payload mass falls within the 2,000-4,000 kg range.

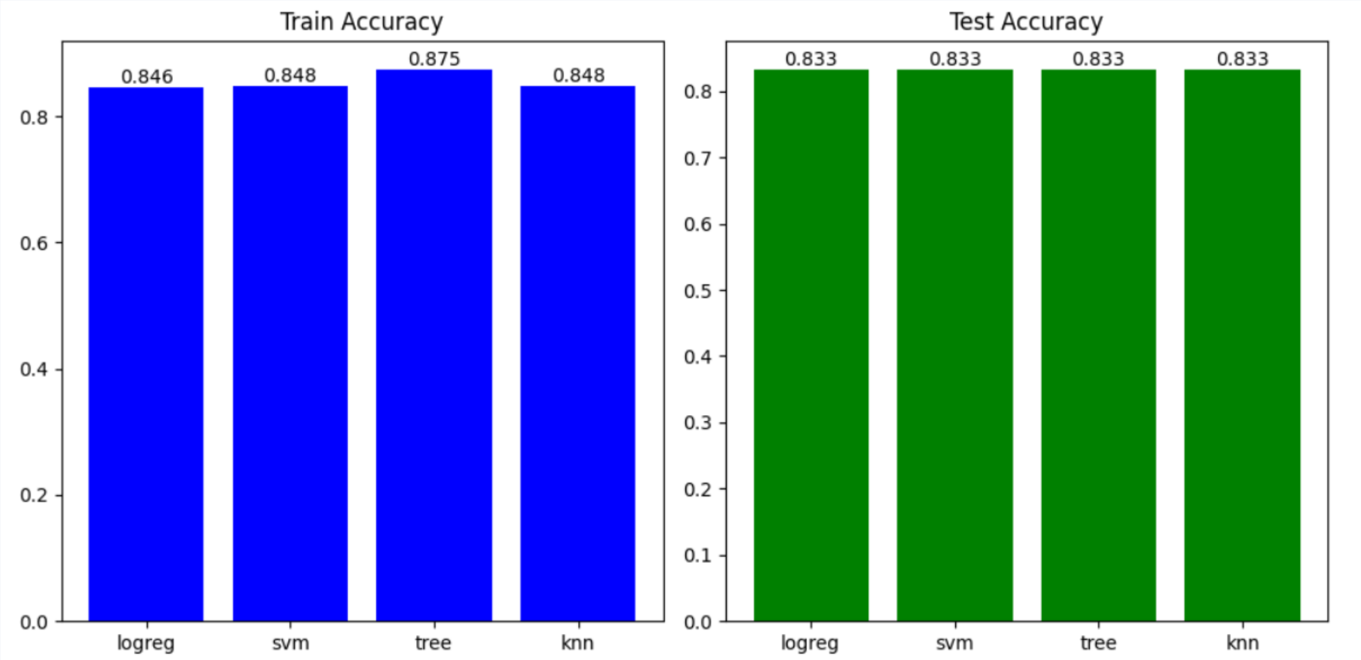


Section 5

Predictive Analysis (Classification)

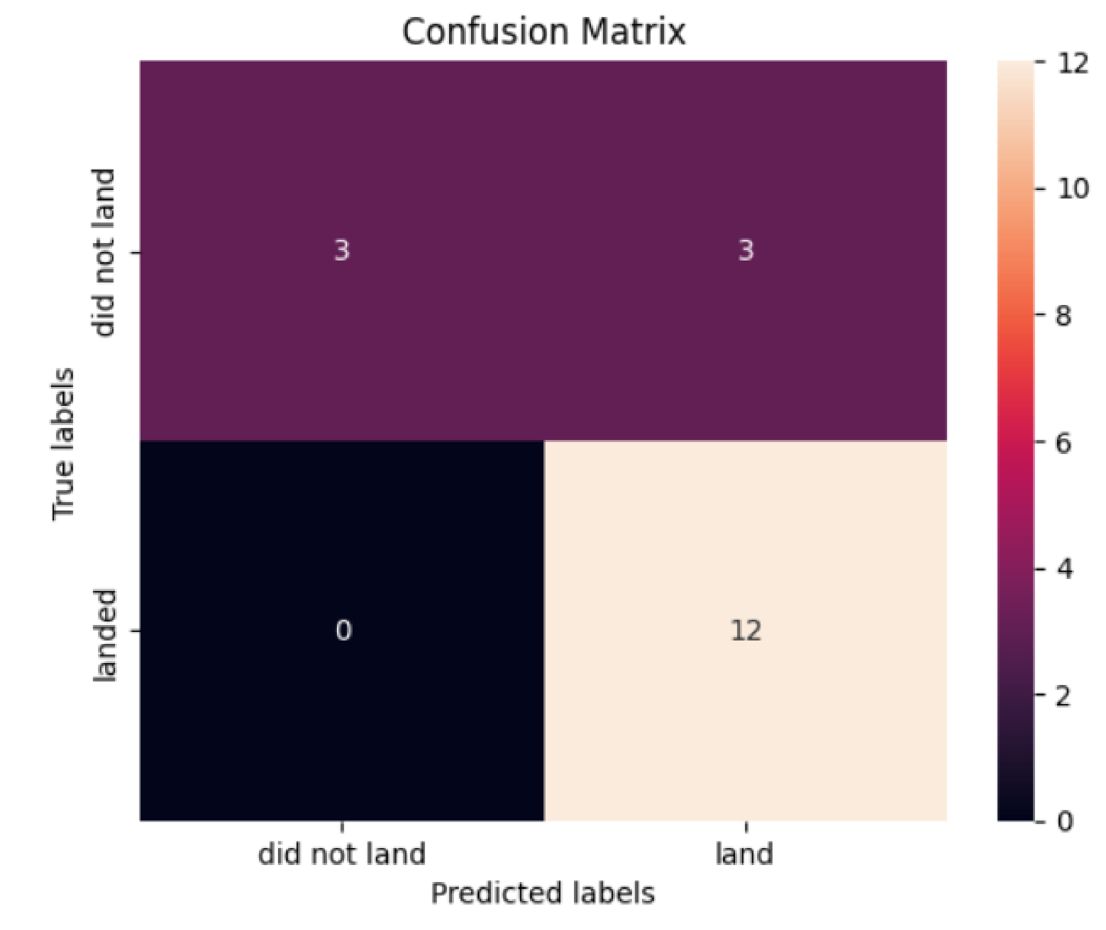
Classification Accuracy

The bar graph on the right displays the accuracy of various classification models during training and testing. While all models achieved the same accuracy in testing, the Decision Tree model demonstrated the highest accuracy during training, outperforming the other models.



Confusion Matrix

The confusion matrix shows the performance of the top-performing Decision Tree model. Notably, the model perfectly classified all instances labeled as "landed" in the true labels, achieving 100% accuracy for this class. However, the model exhibits a bias towards classifying data points as "land" more frequently than "did not land", indicating potential overprediction of successful landings.



Conclusions

- High Payload Mass has a positive relationship with launch-landed success rates
- SpaceX launch-landed success rates have been growing since 2013
- SpaceX launch sites were constructed in coastal areas with important elements such as coastlines, railways, and highways nearby, and cities further away
- KSC LC-39A is the preferable site for rocket launch and landing
- Decision Tree model worked the best in this project
- Decision Tree outperformed Logistic Regression, SVM, and KNN models

Thank you!