IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akhil Nishad
12-June-2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- **Project background and context**

SpaceX reduces launch costs by reusing the Falcon 9 first stage. Predicting whether the first stage will land helps estimate launch costs. In this project, as a data scientist at fictional company SpaceY, I use public SpaceX data and machine learning to predict landings and support business decisions through dashboards.

- **Problems you want to find answers**
  - Can we predict if the first stage will land?
  - What factors affect landing success?
  - How does launch performance vary over time?
  - How can we visualize insights for decision-making?

Section
1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX API and web scraping from wikipedia
- Perform data wrangling
  - One hot encoding was applied to categorical features
- 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

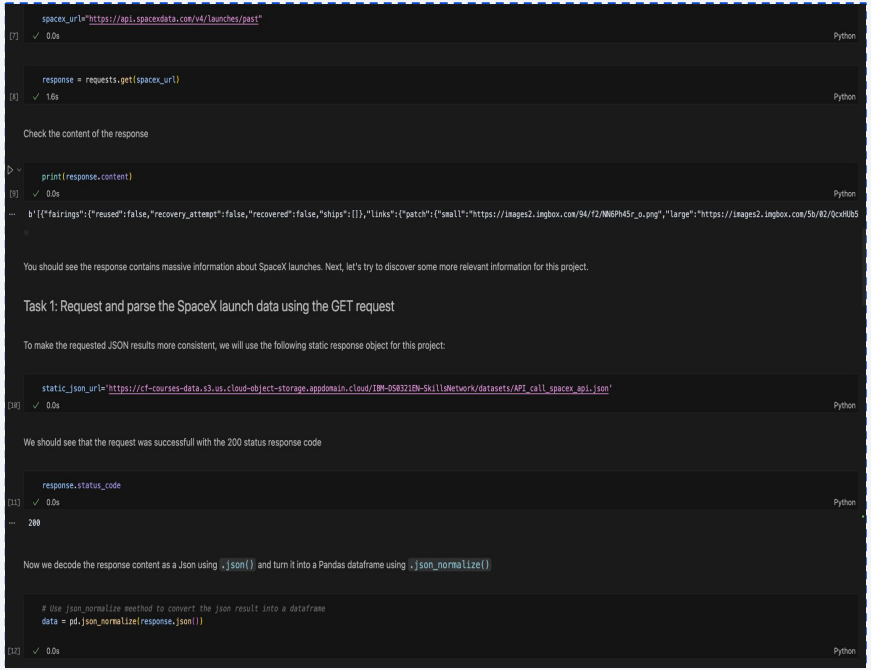
Data Collection

- The data was collected using various methods
  - Data collection was done using get request to the SpaceX API.
  - Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
  - We then cleaned the data, checked for missing values and fill in missing values where necessary.
  - In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
  - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

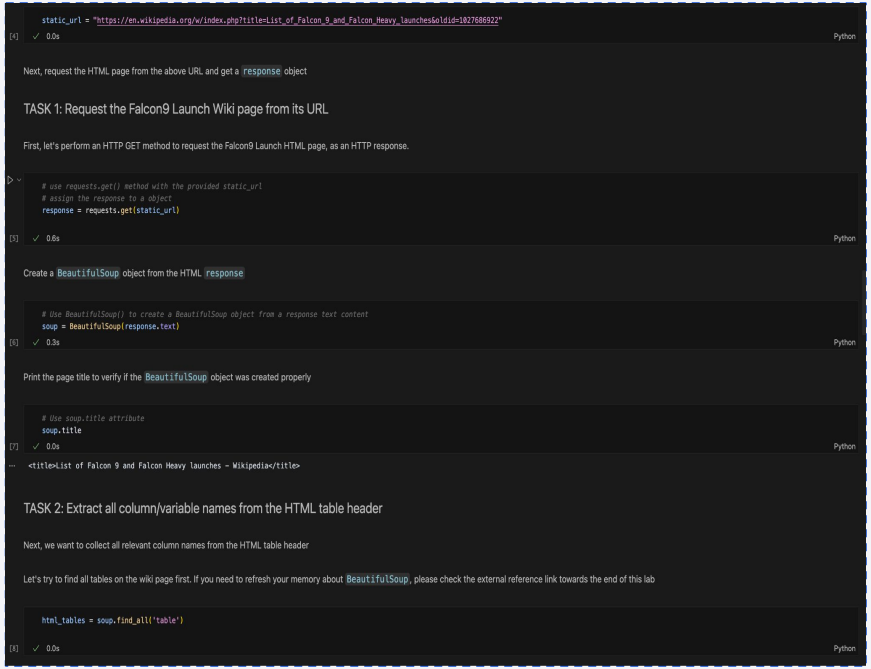
- We retrieved data from the SpaceX API using a GET request, then cleaned, formatted, and performed basic wrangling on the data to prepare it for analysis.

- Github url: [link]



Data Collection - Scraping

- We used web scraping with BeautifulSoup to extract Falcon 9 launch records from the web. The data was parsed from an HTML table and converted into a pandas DataFrame for analysis.
- Github url: [Link]

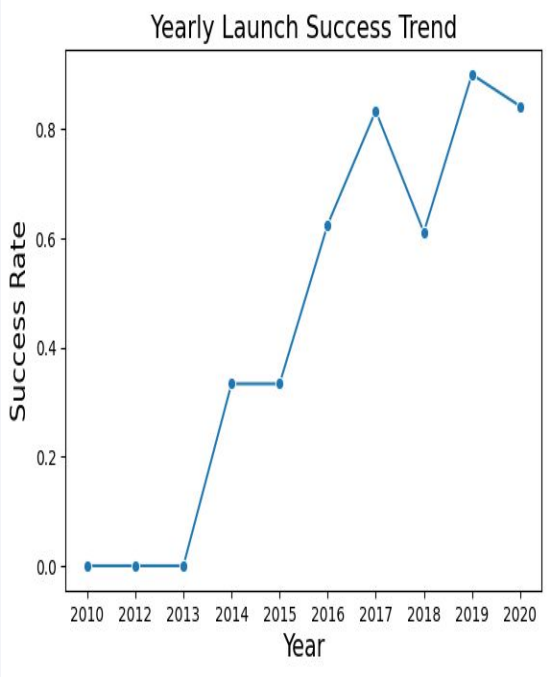


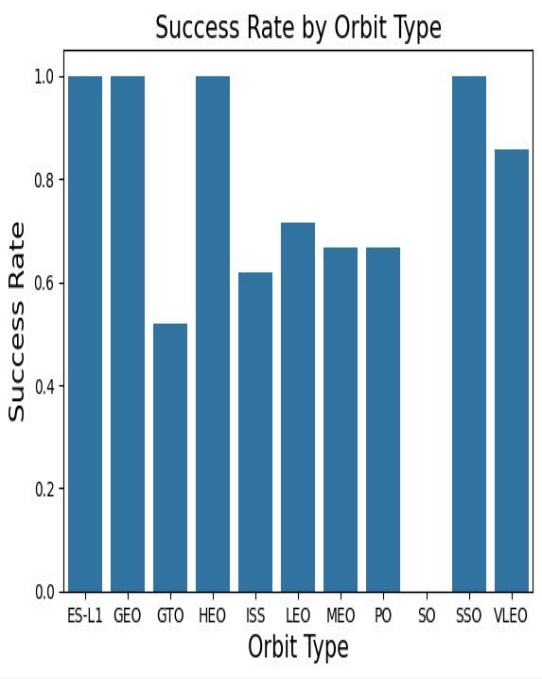
Data Wrangling

- We conducted exploratory data analysis and defined the training labels.
- This included analyzing the number of launches per site and the frequency of different orbit types.
- We also derived the landing outcome label from the outcome column and exported the processed data to a CSV file.
- Github url: Link

EDA with Data Visualization

- We visualized the data to explore relationships such as flight number vs. launch site, payload vs. launch site, success rates by orbit type, flight number vs. orbit type, and launch success trends over the years.
- Github url: Link





EDA with SQL

- We loaded the SpaceX dataset into a SQLite database directly within the Jupyter notebook. Using SQL, we performed exploratory analysis to uncover insights such as:
  - Identifying unique launch site names
  - Calculating total payload mass for NASA (CRS) missions
  - Finding the average payload mass for booster version F9 v1.1
  - Counting total successful and failed mission outcomes
  - Retrieving details of failed drone ship landings, including booster versions and launch sites.

- Github url: Link

Build an Interactive Map with Folium

- We visualized all launch sites on a Folium map, adding markers, circles, and lines to indicate launch outcomes.
- Launch outcomes were classified as 0 for failure and 1 for success.
- By using color-coded marker clusters, we identified launch sites with higher success rates.
- Additionally, we calculated distances from each launch site to nearby locations.
- Github url: Link

Build a Dashboard with Plotly Dash

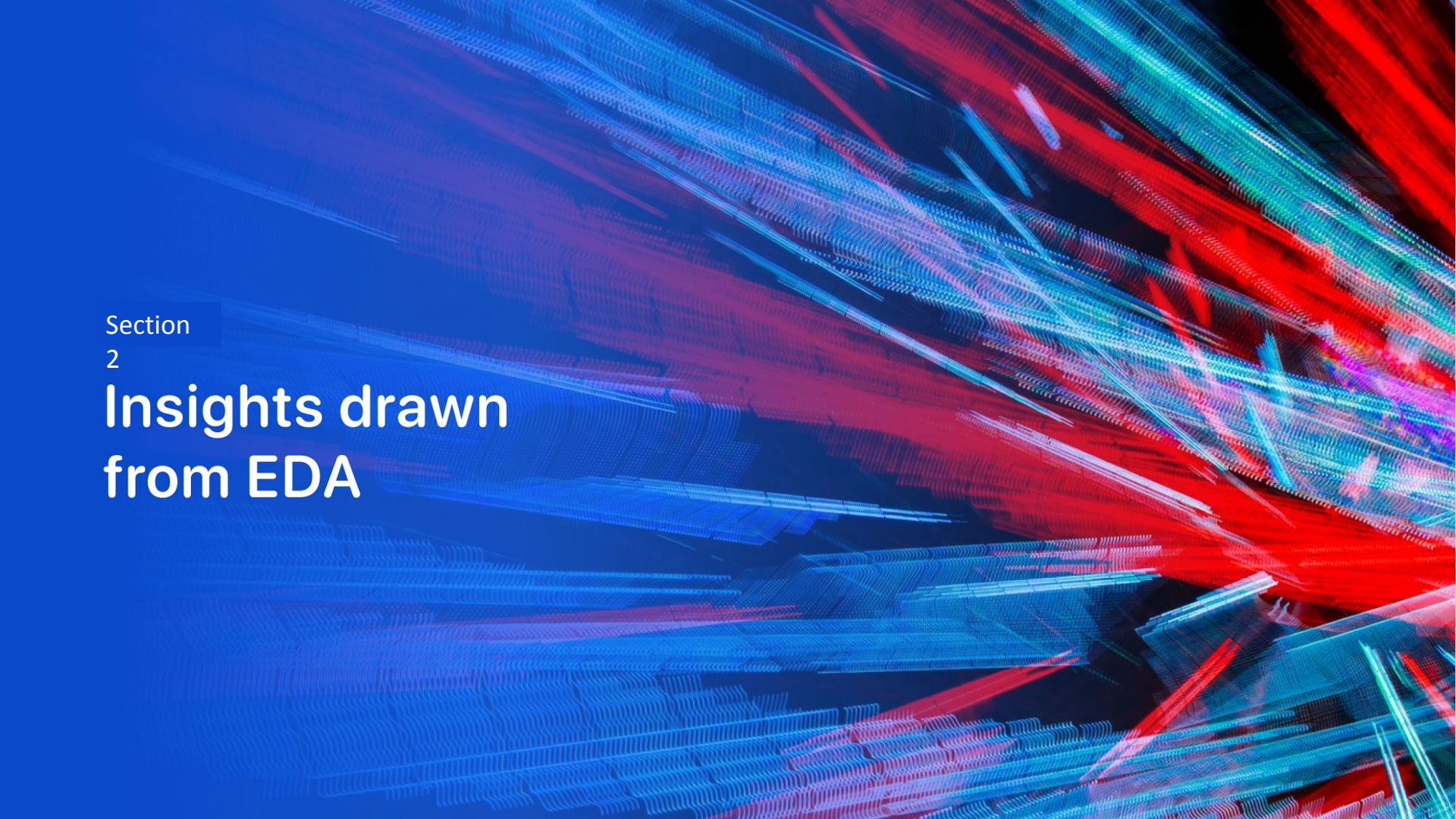
- We developed an interactive dashboard using Plotly Dash.
- It includes pie charts displaying the total number of launches by site, and scatter plots illustrating the relationship between launch outcomes and payload mass (kg) across different booster versions.
- Github url: Link

Predictive Analysis (Classification)

- We used NumPy and pandas to load and preprocess the data, then split it into training and testing sets.
- Multiple machine learning models were built and optimized using GridSearchCV for hyperparameter tuning.
- Accuracy was used as the evaluation metric, and model performance was enhanced through feature engineering and algorithm optimization.
- Finally, we identified the best-performing classification model.
- Github url: Link

Results

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

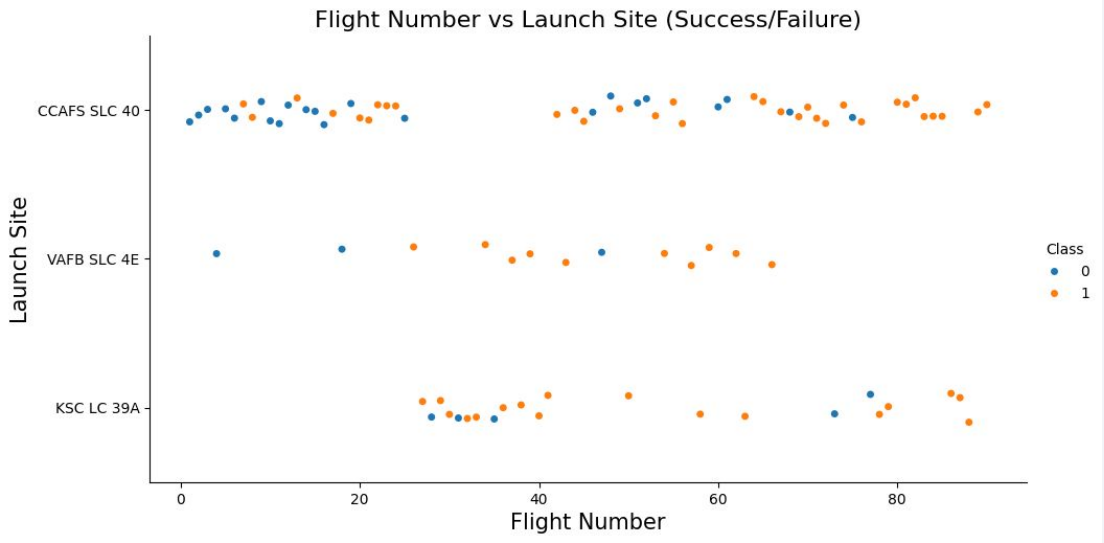


Section 2

Insights drawn from EDA

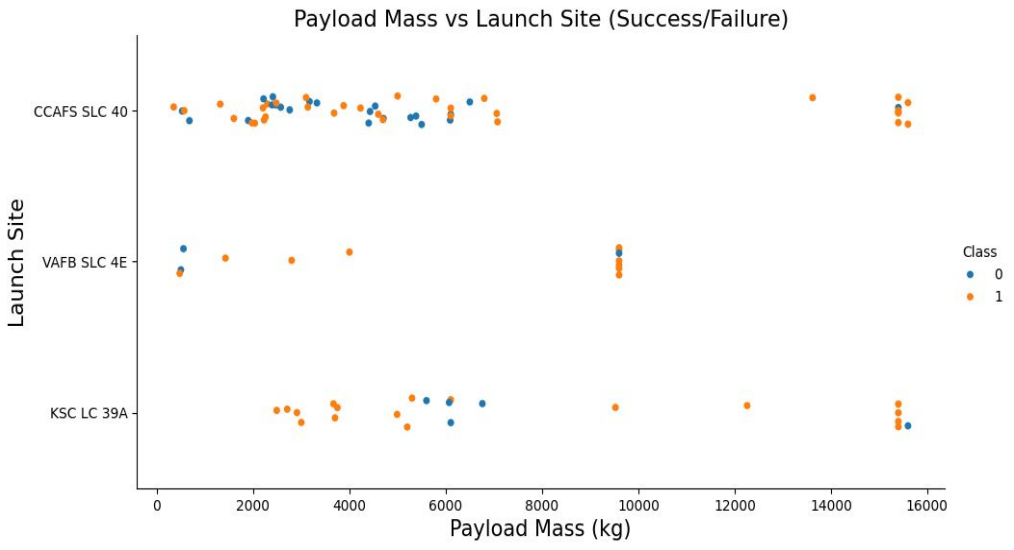
Flight Number vs. Launch Site

- The plot showed that launch sites with a higher number of flights tend to have higher success rates.



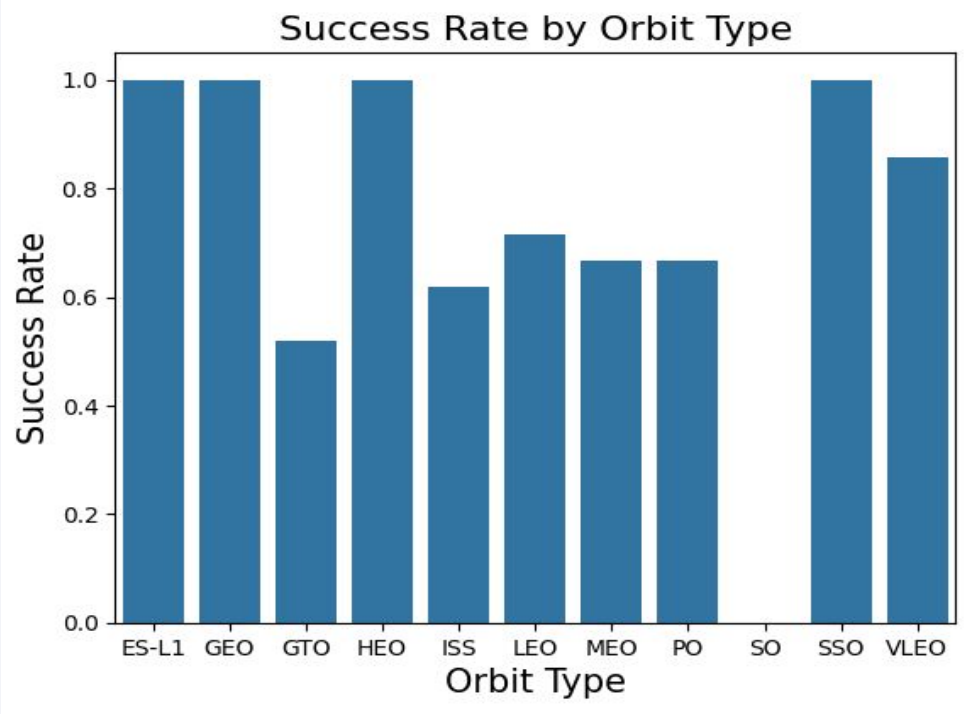
Payload vs. Launch Site

- CCAFS SLC 40 demonstrates the greatest payload versatility, successfully launching missions ranging from ~1,000 kg to over 15,000 kg, while VAFB SLC 4E and KSC LC 39A operate within more constrained payload mass ranges.



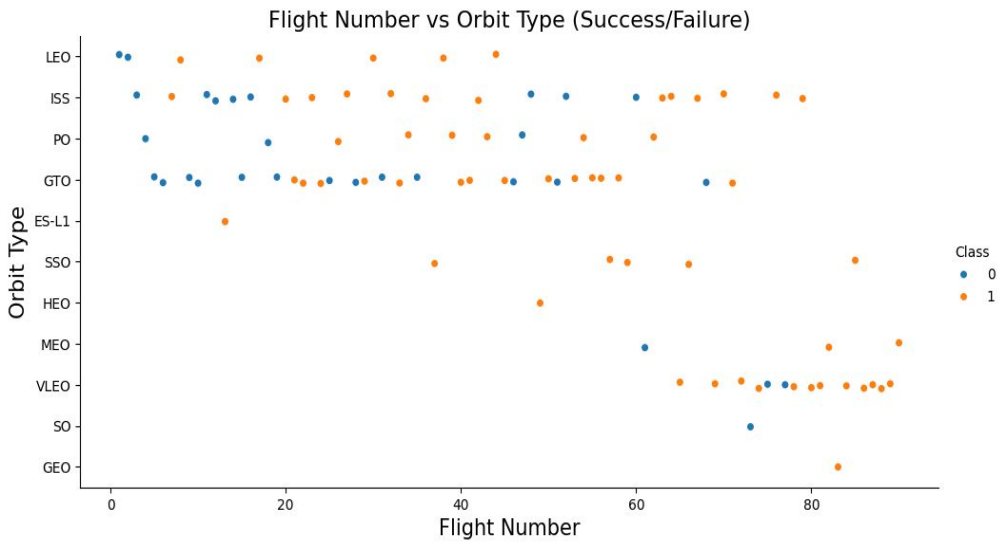
Success Rate vs. Orbit Type

- Mission success rates vary significantly by orbit type, with ES-L1, GEO, HEO, and SSO achieving near-perfect reliability (~100%) while GTO and ISS missions show notably lower success rates (~52-62%).



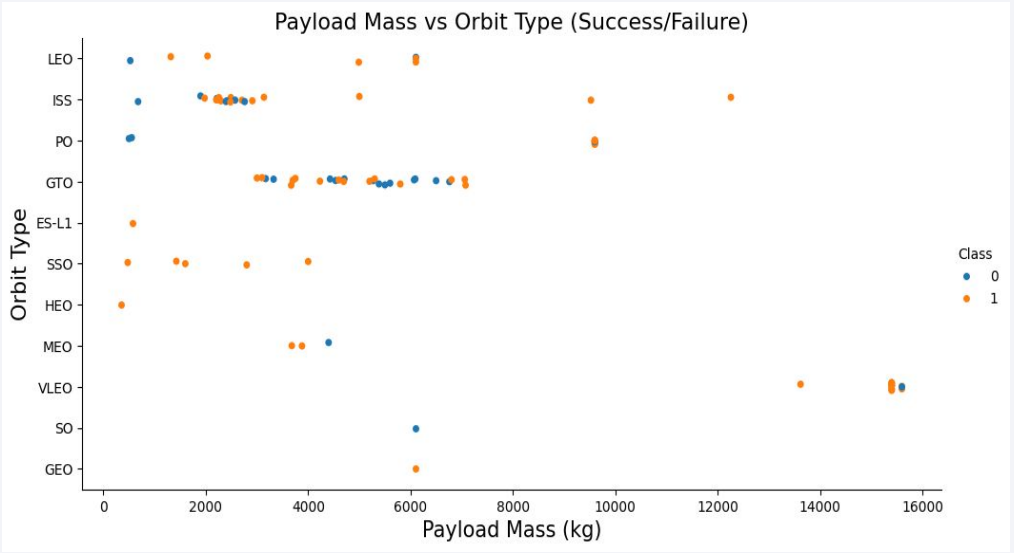
Flight Number vs. Orbit Type

- Mission complexity and frequency patterns emerge over time, with GTO and ISS missions dominating early flights, while specialized orbits like VLEO and SSO become more prominent in later flight numbers, suggesting operational maturity and expanded mission capabilities.



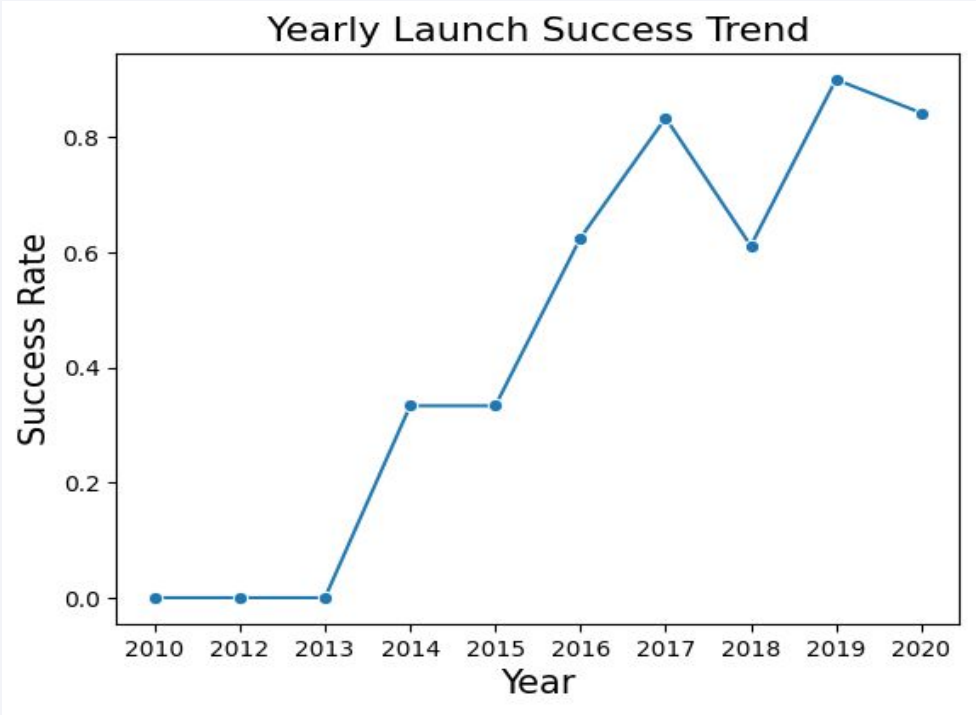
Payload vs. Orbit Type

- Different orbit types accommodate distinct payload mass ranges, with ISS missions clustering around 2,000-3,000 kg, GTO missions spanning 3,000-7,000 kg, and VLEO missions handling the heaviest payloads at 14,000-15,000 kg, reflecting specific mission requirements and orbital mechanics constraints.



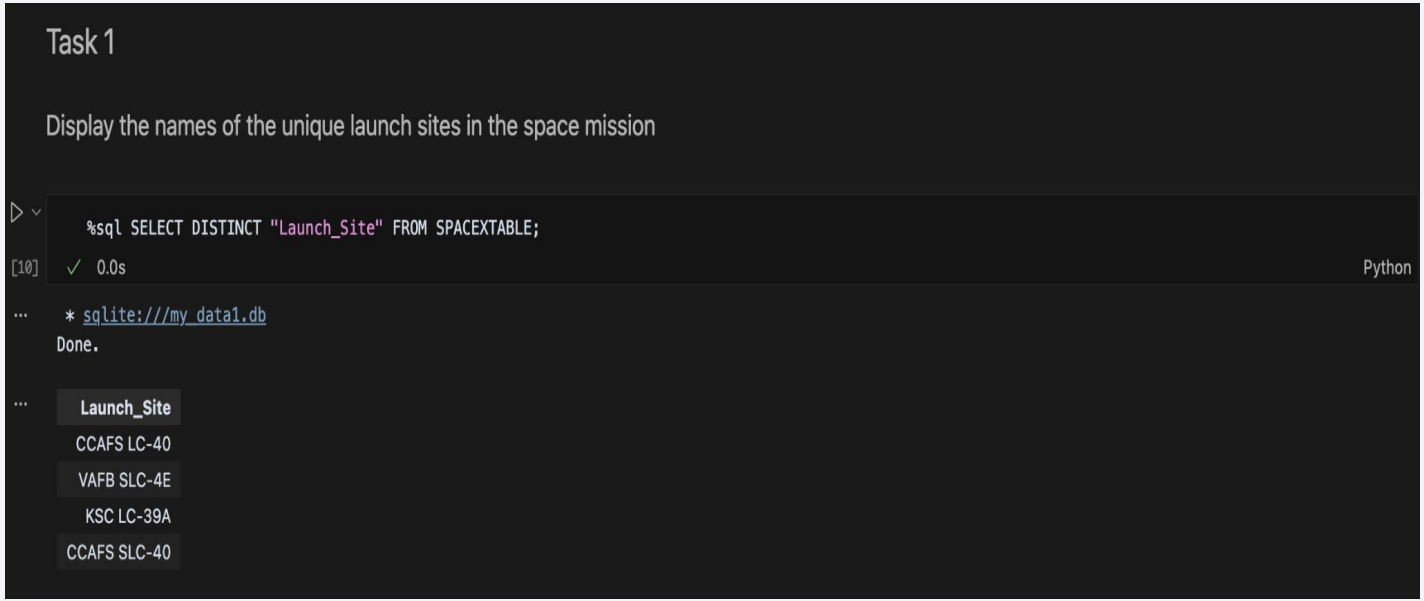
Launch Success Yearly Trend

- Launch success rates demonstrate steady improvement over time, evolving from 0% in early years (2010-2013) to consistently high performance above 80% by 2019-2020, indicating significant operational learning and system maturation.



All Launch Site Names

- We used the keyword DISTINCT to filter out the unique launch site names.



Launch Site Names Begin with 'CCA'

- We used the LIMIT keyword set to 5 to retrieve the first five records of launch sites whose names start with 'CCA'.

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

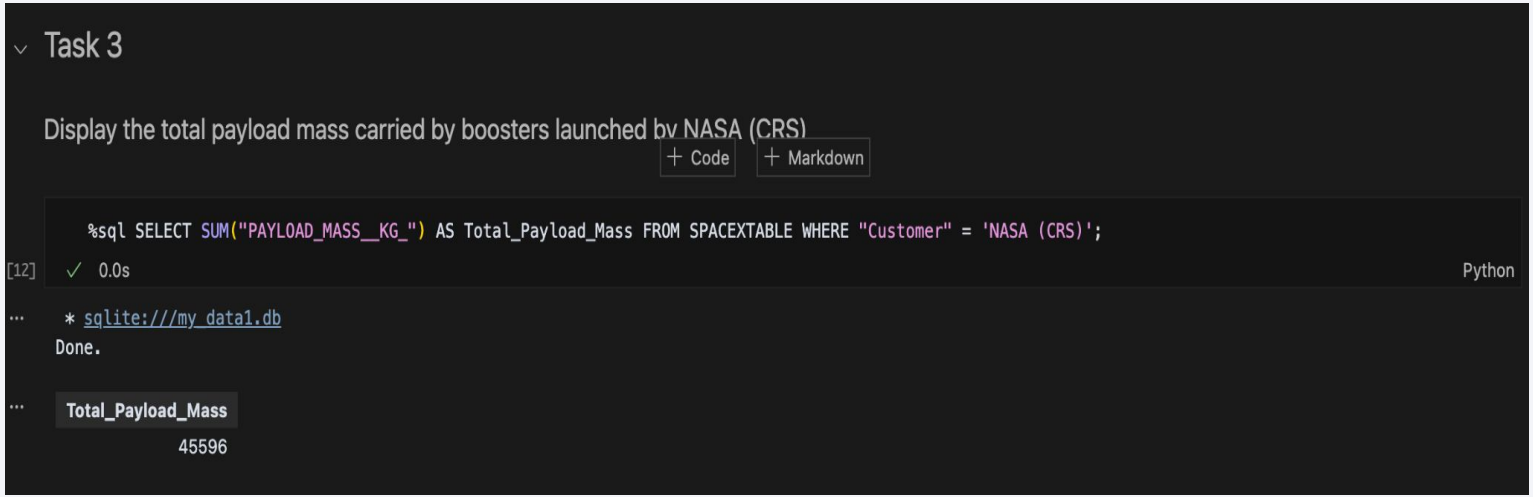
[11] ✓ 0.0s Python

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

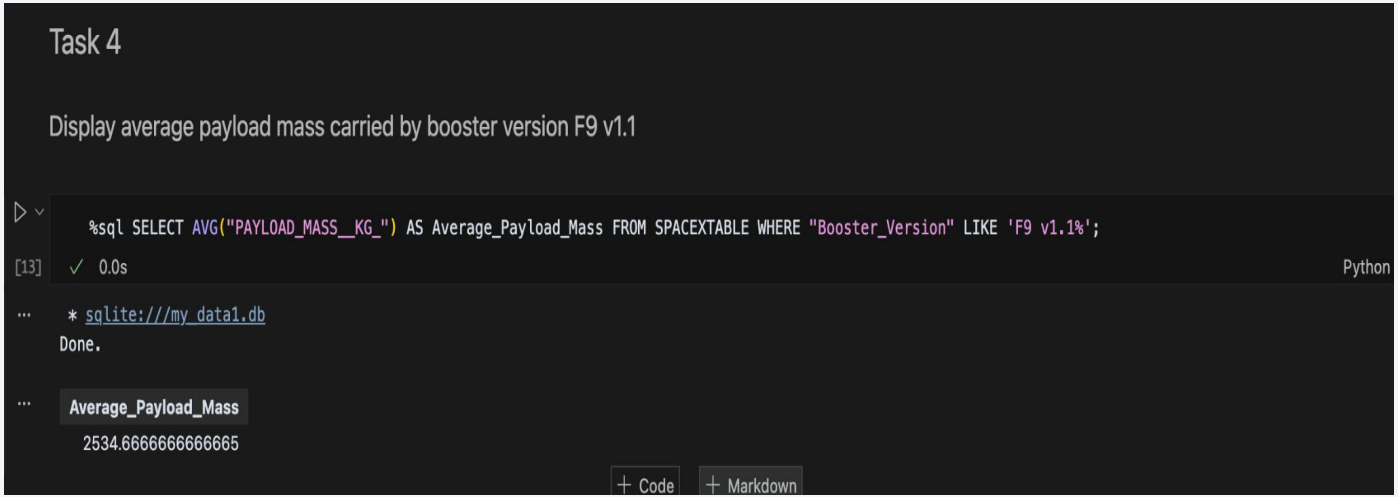
Total Payload Mass

- Using the query below, we determined that the total payload carried by NASA boosters is 45,596.



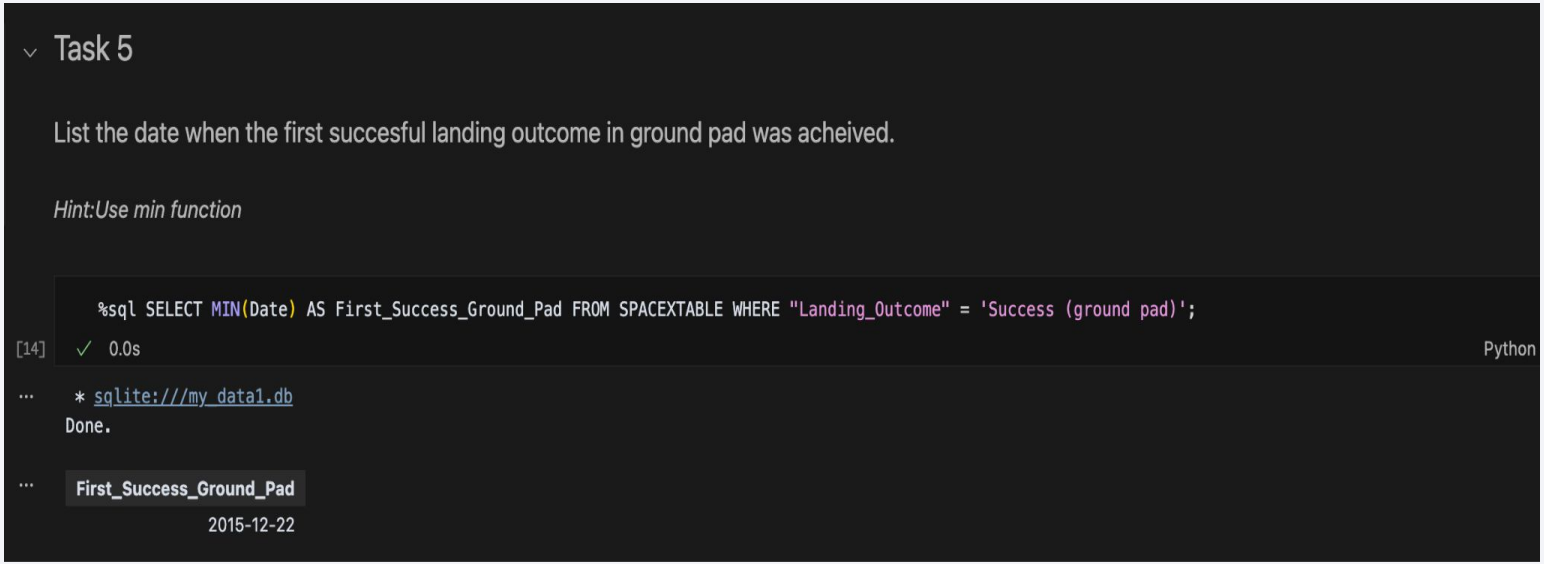
Average Payload Mass by F9 v1.1

- We found that the average payload mass for the F9 v1.1 booster version is 2,534.67 kg.



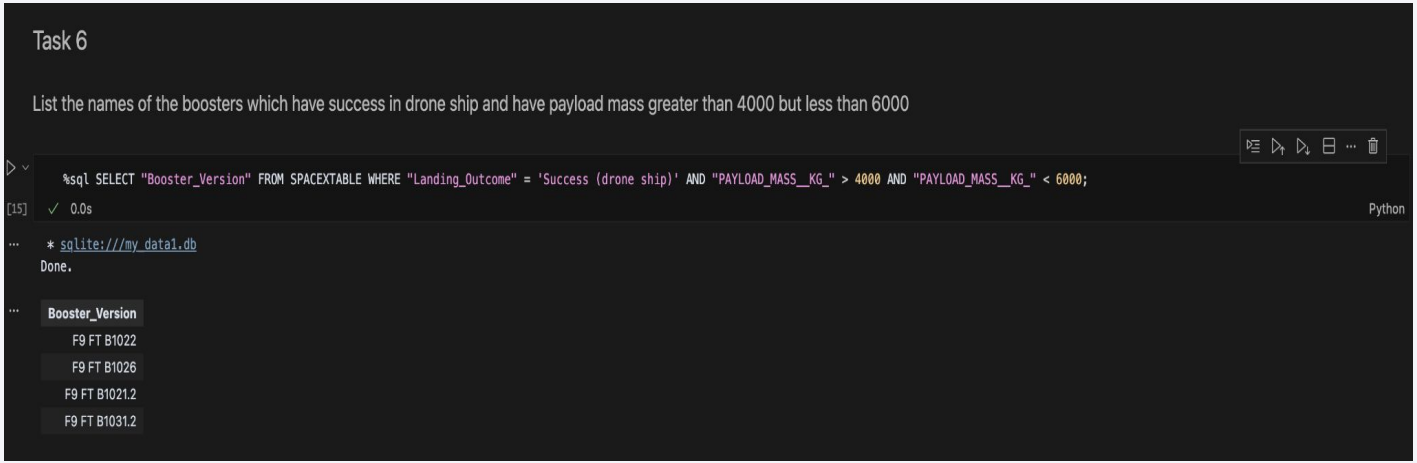
First Successful Ground Landing Date

- The first successful ground landing date was: 22 December 2015.



Successful Drone Ship Landing with Payload between 4000 and 6000

- We applied the WHERE clause to filter boosters that successfully landed on a drone ship, and used the AND condition to select those with a payload mass between 4,000 and 6,000 kg.



Total Number of Successful and Failure Mission Outcomes

- This query counts the number of occurrences for each unique mission outcome, providing a summary of how many launches resulted in success, failure, or other outcomes.

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS Outcome_Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

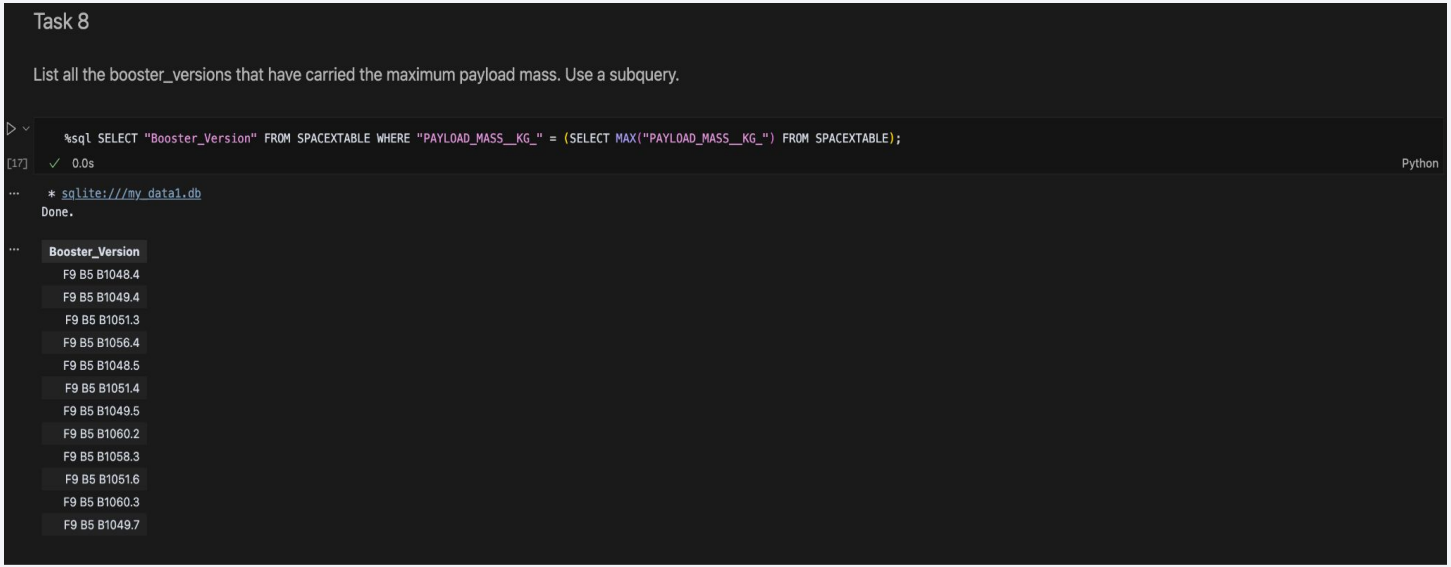
[16] ✓ 0.0s Python

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

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

Boosters Carried Maximum Payload

- We identified the booster that carried the maximum payload by using the MAX() function along with a subquery in the WHERE clause.



2015 Launch Records

- We combined the WHERE clause with LIKE, AND, and BETWEEN conditions to filter records of failed drone ship landings in 2015, retrieving the corresponding booster versions and launch site names.

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date, 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr(Date, 1, 4) = '2015';
```

[18] ✓ 0.0s Python

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

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

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

- We queried the landing outcomes and their counts, filtering records between 2010-03-20 and 2010-06-04 using the WHERE clause. The results were grouped by landing outcome using GROUP BY and sorted in descending order with the ORDER BY clause.

Task 10

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.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Outcome_Count DESC;
```

[19] ✓ 0.0s Python

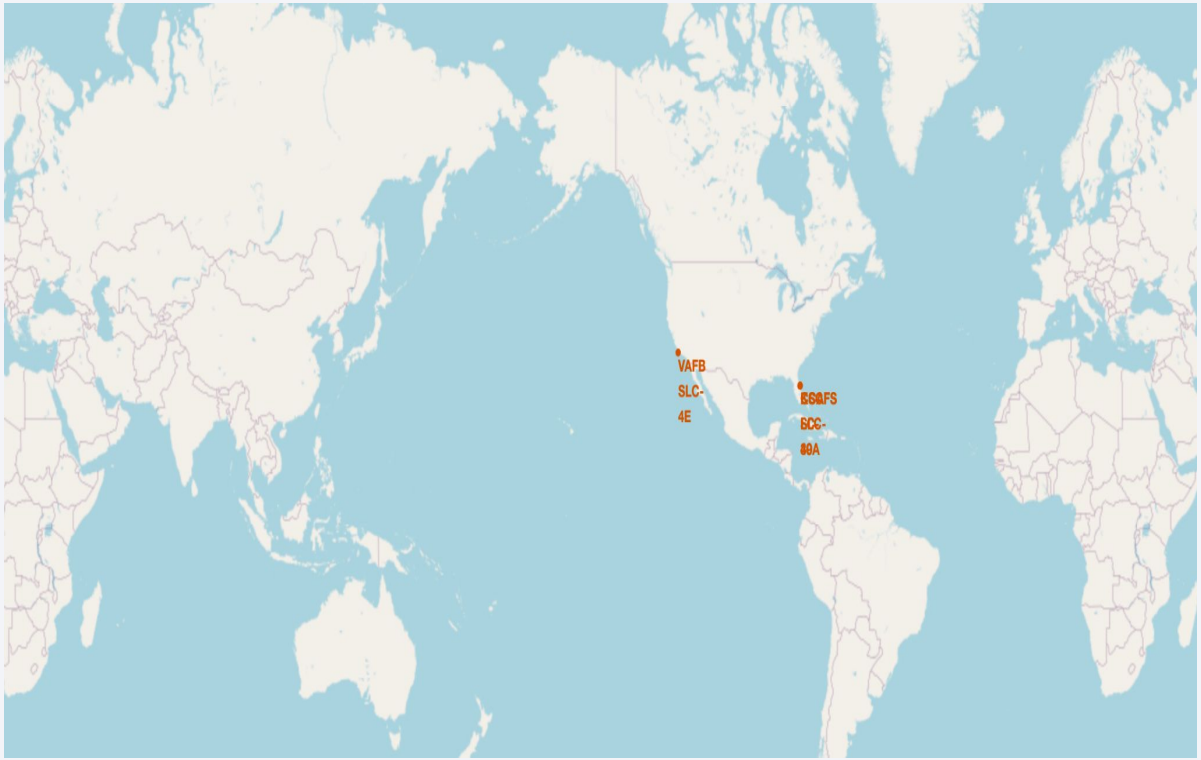
```
* sqlite:///my_data1.db
Done.
```

Landing_Outcome	Outcome_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

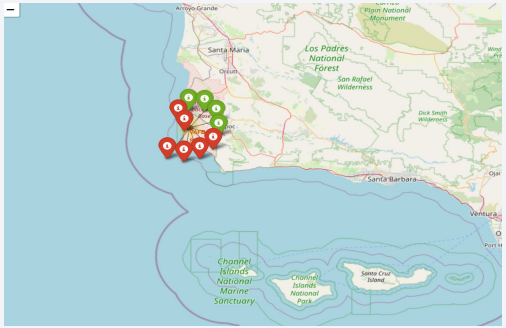
All Launch Sites global map markers

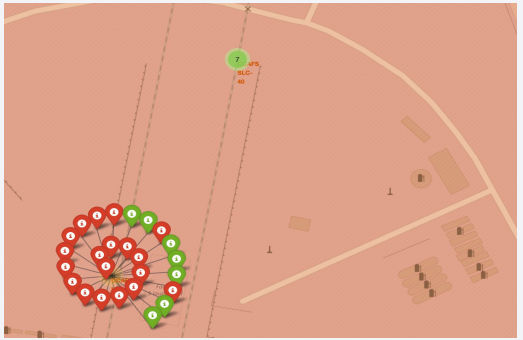


- The launch sites for spacex are along the borders.

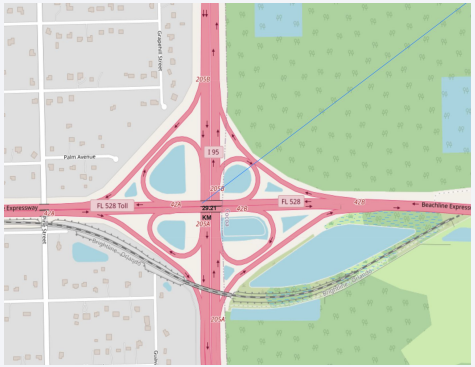
Map showing launch sites with successes and failures

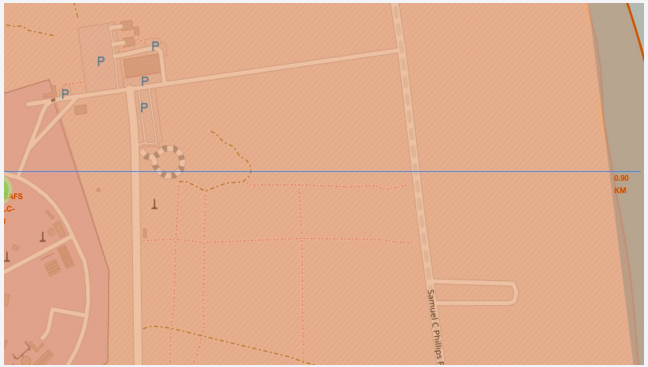
- The green markers show success and the red ones show failure.

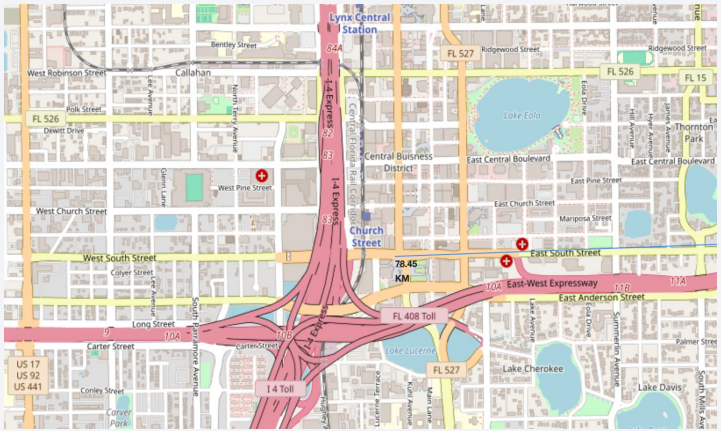




Launch Site distance to Landmarks

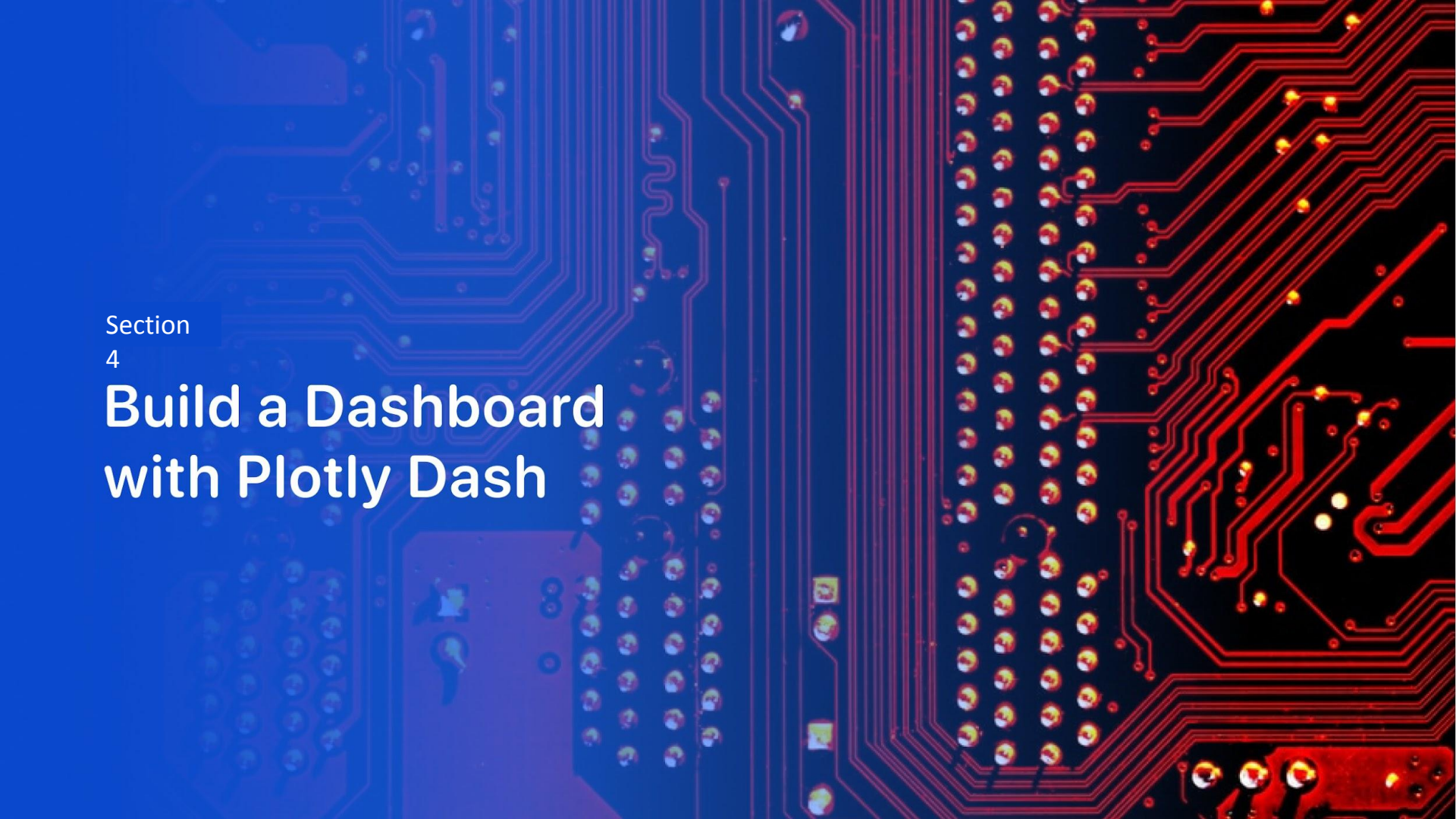




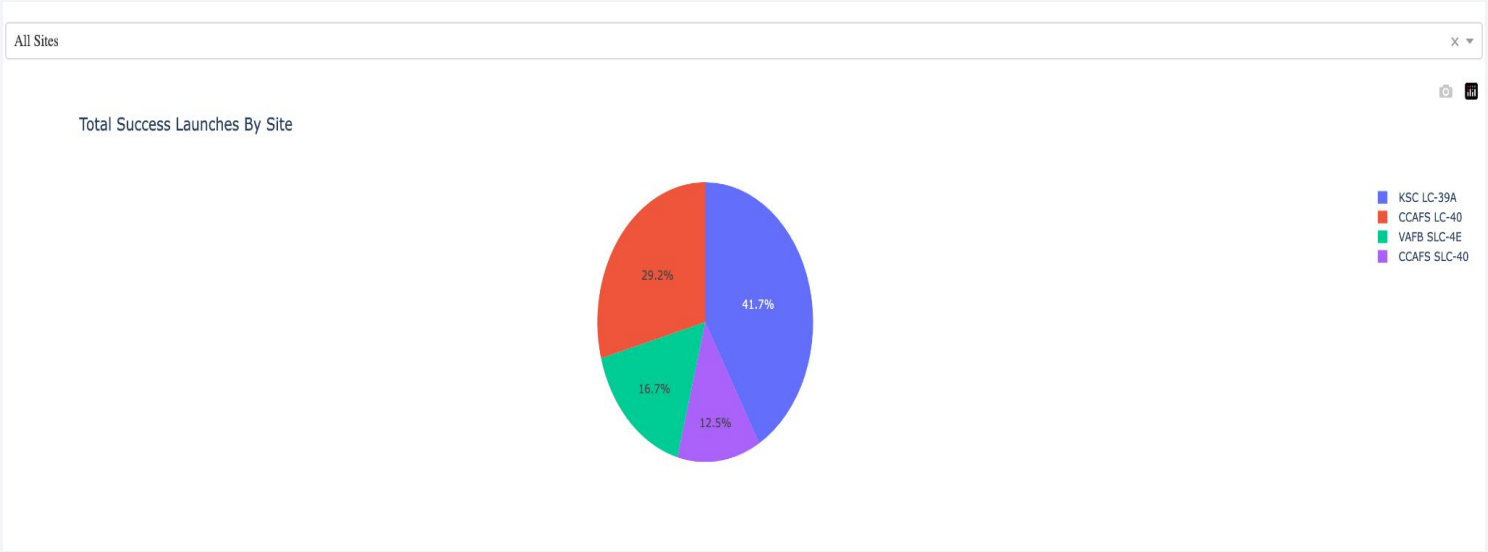


The launch sites are near the cost and are away from major cities. They are also not in close proximities to the highway or railways.

Coast: 0.9km; City: 78.45km; highway: 29.21km

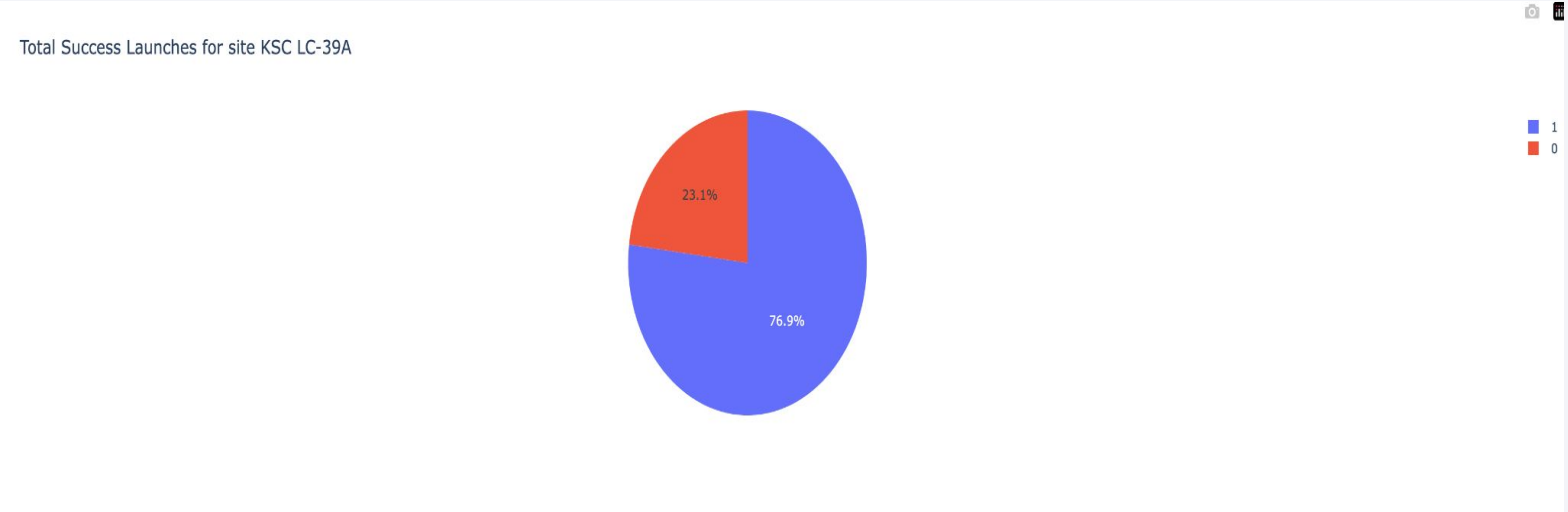

Section
4
Build a Dashboard
with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



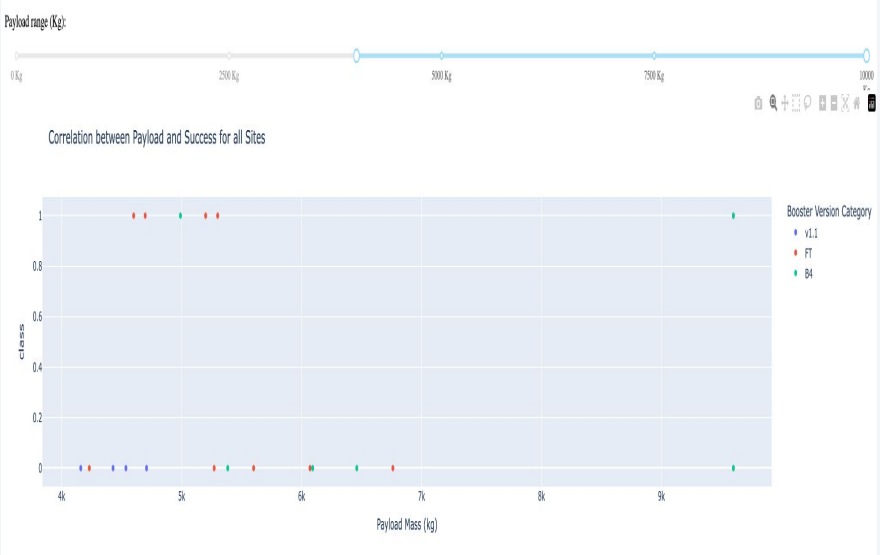
It can be noted that launch site kSC LC-39A had the highest success percentage.

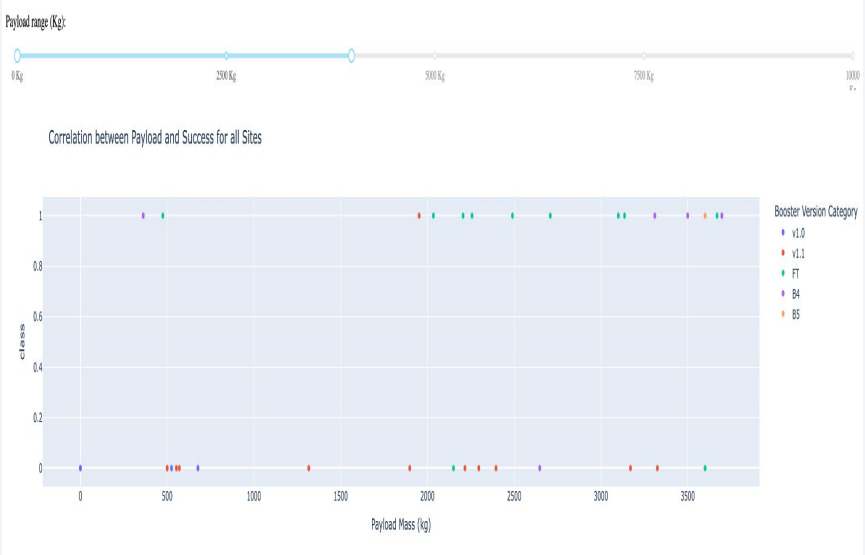
Pie chart showing the Launch site with the highest launch success ratio



The KSC LC-39A achieved a success percentage is 76.9% while the failure is 23.1%.

Scatter plot of Payload vs Launch Outcome for all sites





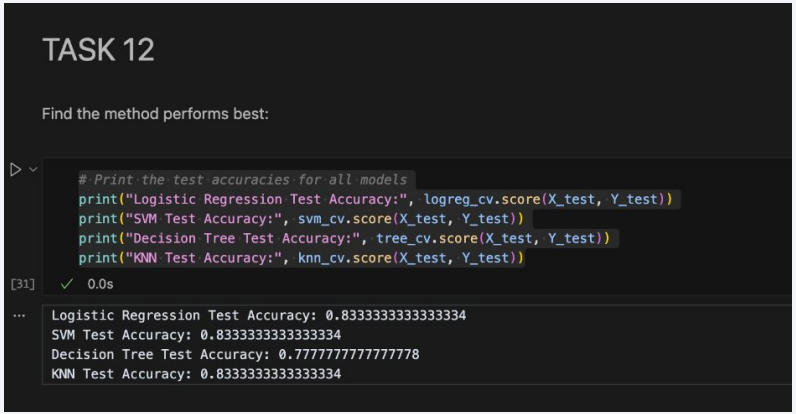
Success for low weighted payloads is greater than that of heavier payloads

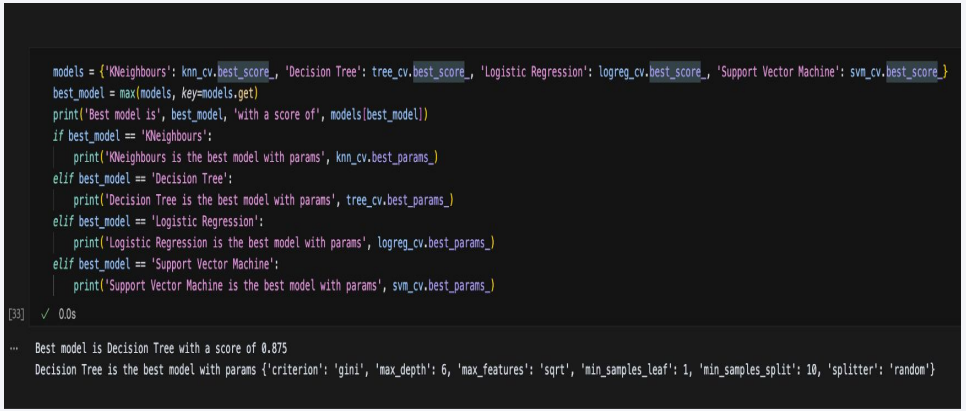
Section
5

Predictive Analysis (Classification)

Classification Accuracy

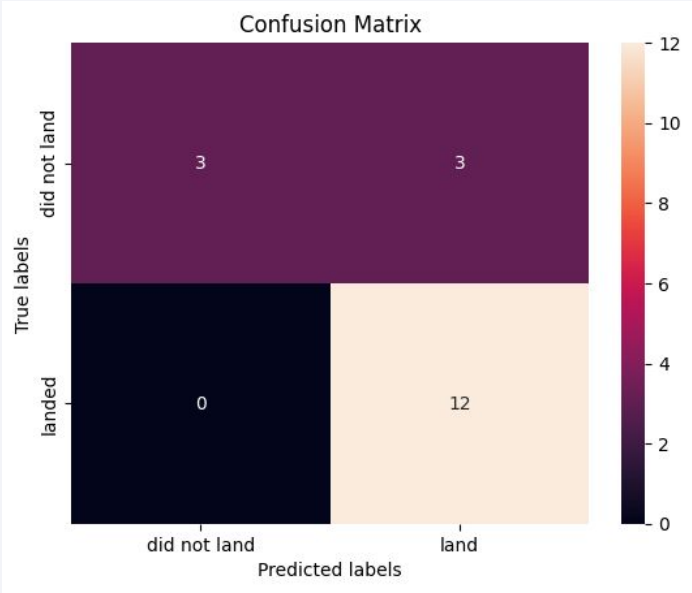
- For the test set, we saw that Logistic Regression, SVM, and KNN all tie for the best accuracy at 83.33% while the decision tree had an accuracy of 77.78%.
- This maybe due to the small test set size. The scores of the entire data set shows that the decision tree is the best model. This model not only has high scores but high accuracy as well.





Confusion Matrix

- The identical performance and matching confusion matrices of the three models may be due to the limited size of the training dataset, which contained only 18 samples, making it difficult to distinguish their effectiveness.



Conclusions

We can conclude the following:

- Launch sites with a higher number of flights tend to have higher success rates.
- Launch success rates began improving steadily from 2013 through 2020.
- Orbits such as ES-L1, GEO, HEO, SSO, and VLEO showed the highest success rates.
- KSC LC-39A recorded the most successful launches among all sites.
- Decision Tree Model is the best algorithm for this dataset.

Thank you!