IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tanushree Dutta
20 September, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection: Utilized the SpaceX REST API and web scraping.
  - Data Wrangling: Created a success/failure outcome variable.
  - Data Exploration: Applied data visualization techniques to examine factors such as payload, launch site, flight number, and yearly trends.
  - Data Analysis: Used SQL to calculate key statistics.
  - Launch Site Analysis: Explored success rates of launch sites in relation to geographical markers.
  - Predictive Modeling: Built Machine Learning Classification models to predict landing outcomes.

- Summary of all results
  - Launch success rates have improved over time, with KSC LC-39A having the highest success rate among landing sites and orbits ES-L1, GEO, HEO, and SSO achieving 100% success.
  - Most launch sites are located near the equator and close to the coast.
  - All predictive models performed similarly on the test set, with the decision tree model showing slightly better performance.

Introduction

- SpaceX rockets are renowned for their reusable technology and successful vertical landings, particularly the Falcon 9.
- Falcon 9 has a high reliability and efficiency track record, although there are occasional crash incidents.
- This Capstone project aims to predict the successful outcomes of Falcon 9 rocket launches using classification models, focusing on the first stage.
- The primary objective is to determine whether the first stage of the rocket will be reused after launch.
- SpaceX's innovative approach to reusing the first stage significantly reduces launch costs, making space travel more affordable.
- The project will explore factors influencing first-stage landing success, as well as analyze the rate of successful landings over time and identify the best predictive model for successful landings.

Section 1

Methodology

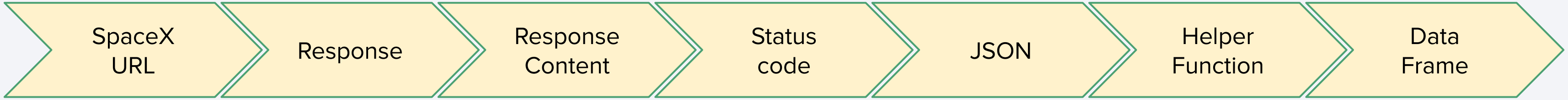
Methodology

Executive Summary

- Collect data using the SpaceX REST API and web scraping techniques.
- Prepare data for analysis by filtering, handling missing values, and applying one-hot encoding.
- Conduct exploratory data analysis (EDA) using SQL and data visualization techniques.
- Visualize data with Folium and Plotly Dash.
- Build and tune classification models to predict landing outcomes, evaluating to find the best model and parameters.

Data Collection - SpaceX API

- ➔ Fetch rocket launch data from the SpaceX API.
- ➔ Decode the response with .json() and transform it into a DataFrame using .json_normalize().
- ➔ Utilize custom functions to extract specific launch information from the SpaceX API.
- ➔ Generate a dictionary from the retrieved data.
- ➔ Create a DataFrame from the dictionary.
- ➔ Filter the DataFrame to retain only Falcon 9 launches.
- ➔ Fill missing values in the Payload Mass column with the mean.
- ➔ Save the final dataset to a CSV file.

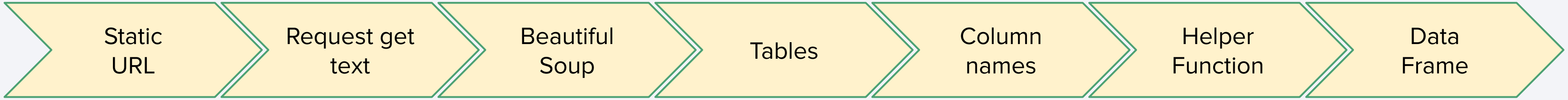


SpaceX URL → Response → Response Content → Status code → JSON → Helper Function → Data Frame

GitHub Link

Data Collection - Scraping

- ➔ Request Falcon 9 launch data from Wikipedia.
- ➔ Create a BeautifulSoup object from the HTML response.
- ➔ Extract column names from the table header in the HTML.
- ➔ Parse the HTML tables to collect data.
- ➔ Convert the parsed data into a dictionary.
- ➔ Build a DataFrame from the dictionary.
- ➔ Export the DataFrame to a CSV file.

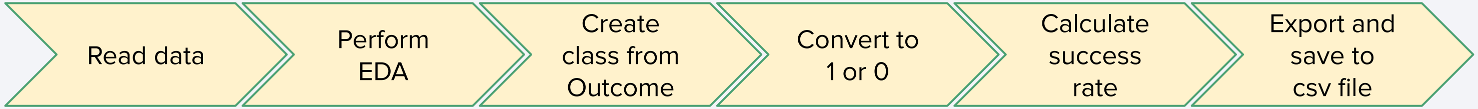


Static URL → Request get text → Beautiful Soup → Tables → Column names → Helper Function → Data Frame

GitHub Link

Data Wrangling

- ➔ The cleaned data was imported and checked for missing values.
- ➔ The data types of all columns were examined.
- ➔ A new feature called "class" was created from the outcome column, where entries labeled "False" or "None" were assigned a value of 0 (bad), and others were assigned 1 (good).
- ➔ The success rate of the "good" outcomes was calculated, accounting for 66.67% of the class feature.



Read data

Perform EDA

Create class from Outcome

Convert to 1 or 0

Calculate success rate

Export and save to csv file

GitHub Link

EDA with Data Visualization

- ➔ Scatter plots were used to visualize how variables such as Payload Mass vs. Flight Number and Launch Site vs. Payload affect launch outcomes.
- ➔ Bar charts depicted the success rates across different orbit types, while scatter plots illustrated the relationships between Flight Number, Orbit Type, and Payload.
- ➔ A line plot was utilized to show the yearly trend of launch successes.

GitHub Link

EDA with SQL

Exploratory data analysis using SQL included the following:

- ➔ Identified unique launch sites for space missions.
- ➔ Retrieved five records where launch sites begin with ‘CCA’.
- ➔ Calculated the total payload mass for NASA (CRS) missions.
- ➔ Determined the average payload mass carried by booster version F9 v1.1.
- ➔ Found the first successful landing outcome on a ground pad.
- ➔ Identified successful drone ship landings for payloads between 4000 and 6000 kg.
- ➔ Counted the total number of successful and failed mission outcomes.
- ➔ Found the booster version with the maximum payload mass.
- ➔ Listed the months of failed drone ship landings in 2015.
- ➔ Ranked landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

GitHub Link

Build an Interactive Map with Folium

- ➔ Added a blue circle marker at NASA Johnson Space Center with a popup displaying its name and coordinates.
- ➔ Placed red circle markers at all launch site coordinates, with popups showing their names.
- ➔ Used colored markers to indicate launch outcomes: green for successful launches and red for unsuccessful ones, highlighting site success rates.
- ➔ Added colored lines to represent the distances between CCAFS SLC-40 and nearby proximities, such as the coastline, railway, highway, and city.

GitHub Link

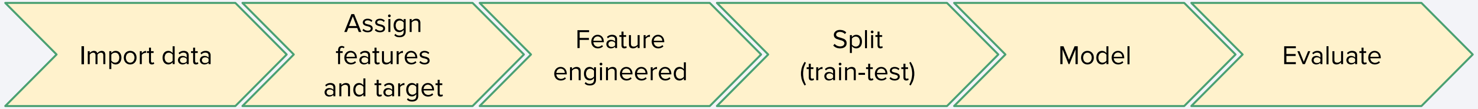
Build a Dashboard with Plotly Dash

- ➔ Dropdown menu enables users to select either all launch sites or a specific launch site.
- ➔ Payload mass slider allows users to filter launches by a selected mass range.
- ➔ Pie chart displays the percentage of successful vs. unsuccessful launches.
- ➔ Scatter plot shows the relationship between payload mass and success rate, categorized by booster version.

GitHub Link

Predictive Analysis (Classification)

- ➔ Loaded data and split it into features and target variables.
- ➔ Normalized feature columns and converted the target column to a NumPy array.
- ➔ Split data into training and test sets.
- ➔ Applied GridSearchCV to find the best parameters and scores for all classification algorithms.
- ➔ Calculated test set accuracy using the .score() method.
- ➔ Plotted a confusion matrix to visualize the model's performance.



GitHub Link

Results

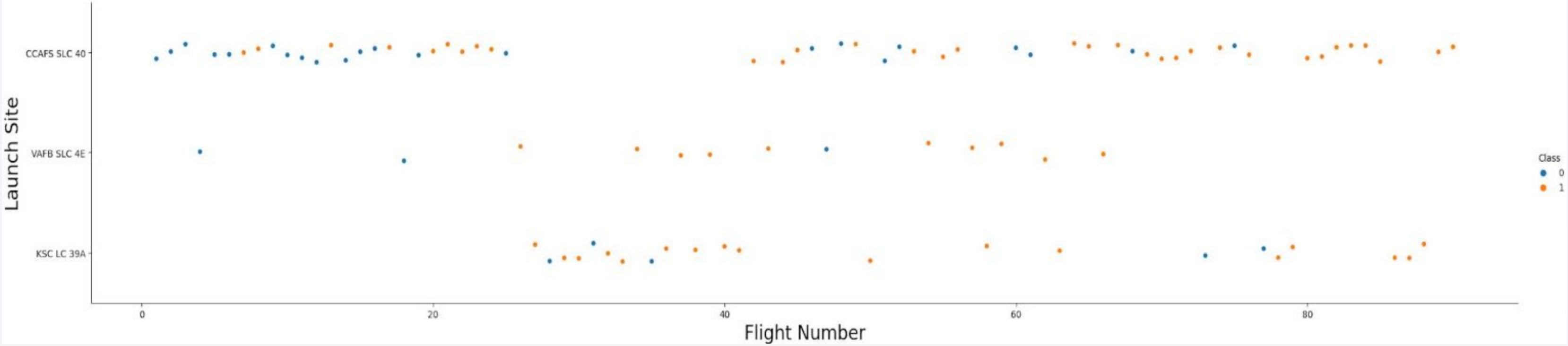
- ➔ NASA (CRS) missions were found to have a total payload mass of 45,596 kg, with the average payload mass for booster version F9 v1.1 being 2928.4 kg. The first successful ground pad landing occurred on 2015-12-22, and five booster versions had successful drone ship landings with payloads between 4000 and 6000 kg. Falcon 9 achieved a total of 99 successful mission outcomes.
- ➔ Visualizations revealed that CCAFS SLC-40 had higher success rates as flight numbers increased, and a 100% success rate was noted for orbits ES-L1, SSO, HEO, and GEO. Falcon 9's first-stage landing success significantly improved between 2010 and 2020.
- ➔ Four machine learning algorithms were used: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and KNN. All models achieved similar accuracy, around 83%.

Section 2

Insights drawn from EDA

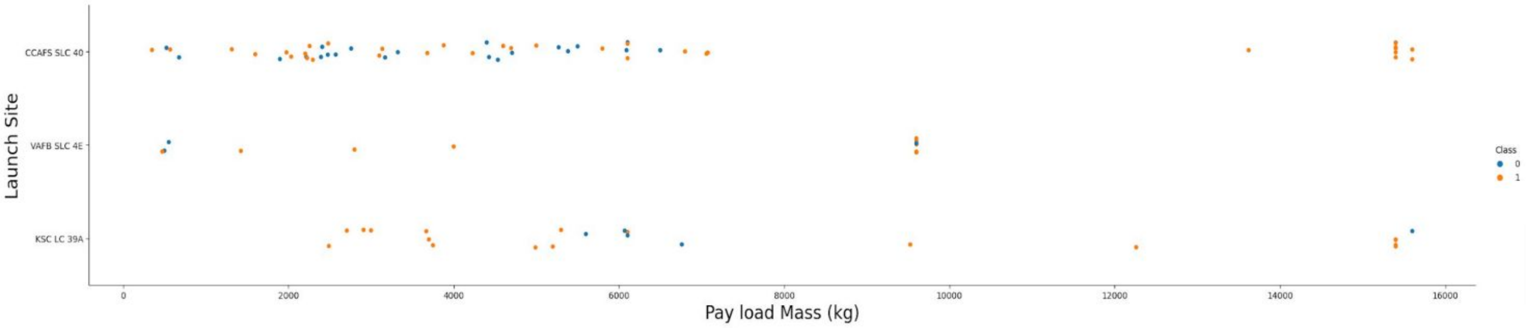
Flight Number vs. Launch Site

- ➔ The analysis showed no success or failure metrics for KSC LC 39A at lower flight numbers, while VAFB SLC 4E had two failures; as flight numbers increased past 80, CCAFS SLC 40 had more successes, and VAFB SLC 4E had none.
- ➔ It was observed that CCAFS SLC 40 launched the highest number of rockets compared to other sites.
- ➔ Additionally, later flights from VAFB SLC 4E and KSC LC 39A exhibited a higher



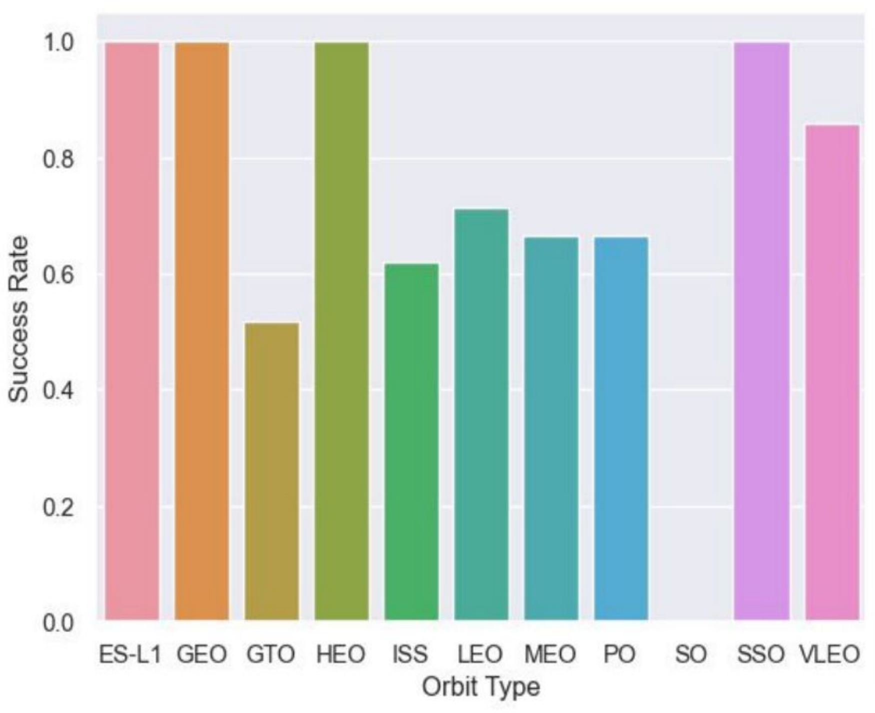
Payload vs. Launch Site

- ➔ The analysis showed that at lower flight numbers (20),VAFB SLC 4E had two failures, and CCAFS SLC 40 had more failures than successes; however, as flight numbers exceeded 80, CCAFS SLC 40 had more successes, while VAFB SLC 4E had none.
- ➔ Most rockets launched across all sites had a payload mass of less than 9,000 kg.
- ➔ CCAFS SLC 40 exhibited a higher success rate for rockets launched with heavy payload masses of 14,000 kg and 16,000 kg compared to VAFB SLC 4E and KSC LC 39A.



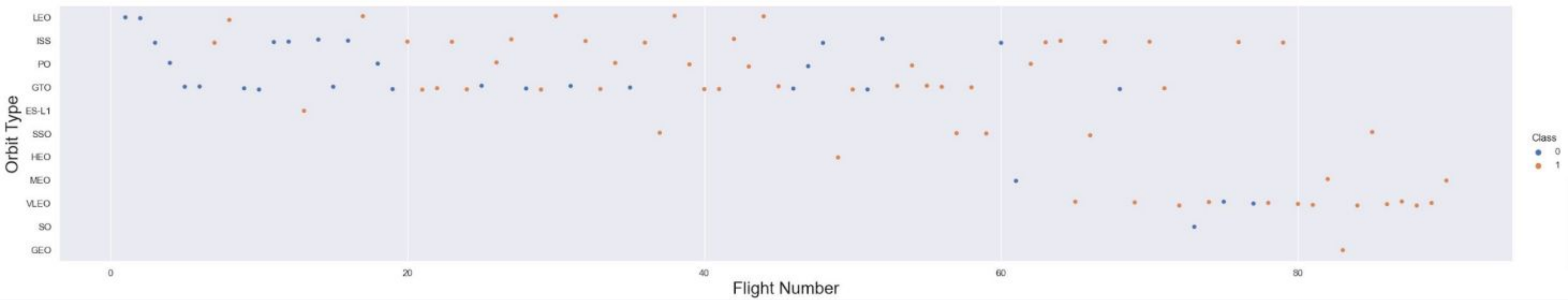
Success Rate vs. Orbit Type

- ➔ Orbits VLEO, ES-L1, GEO, HEO, and SSO have the highest success rates among orbit types.
- ➔ Orbit SO has the lowest success rate.



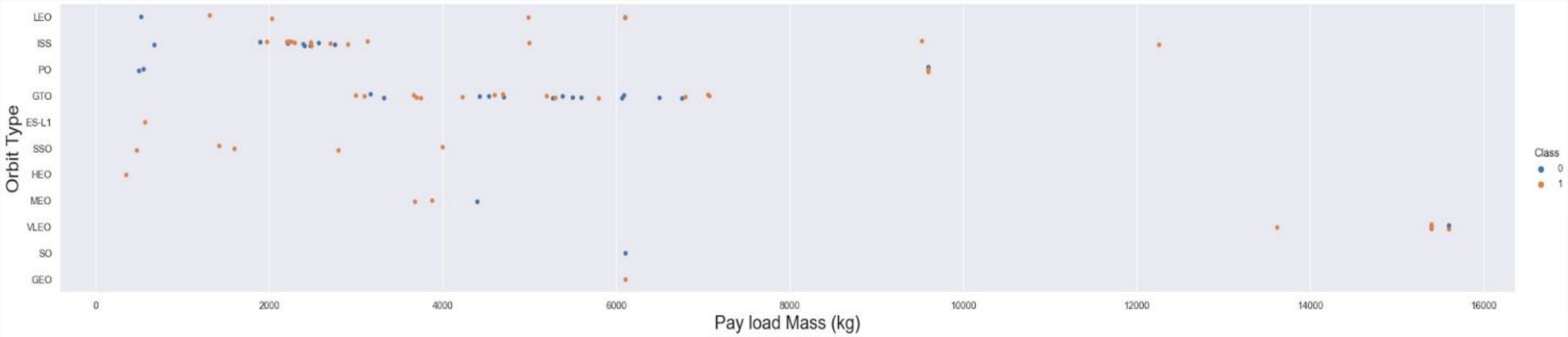
Flight Number vs. Orbit Type

- ➔ More rockets were launched in the LES, ISS, PO, GTO, and VLEO orbits.
- ➔ In the LEO orbit, success appears related to the number of flights, while no relationship was observed between flight number and other orbits.



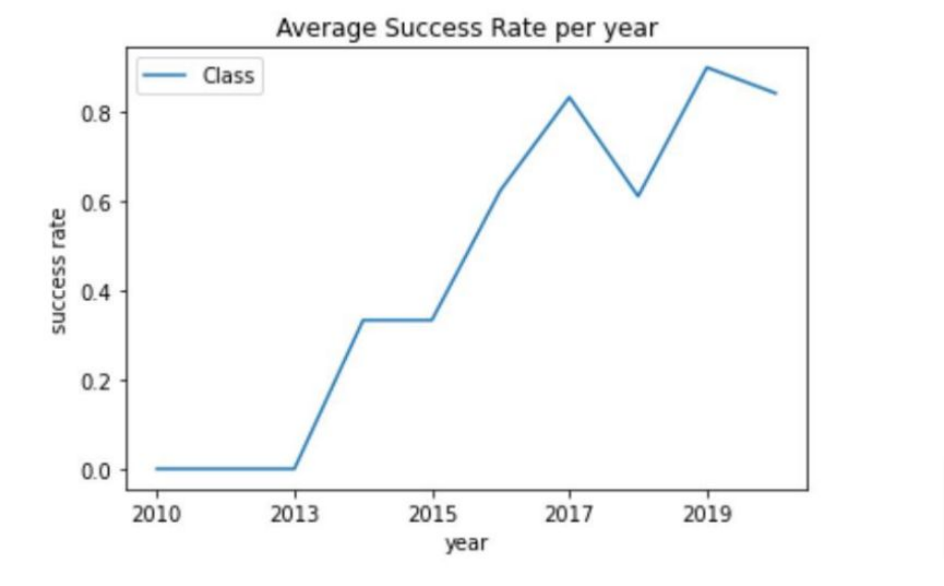
Payload vs. Orbit Type

- ➔ Higher success rates are observed for rockets with heavy payloads launched in PO, LEO, and ISS orbits.
- ➔ SSO and MEO orbits have high success rates for rockets with lighter payloads.
- ➔ GTO rockets show both positive and negative landing rates, regardless of payload size.



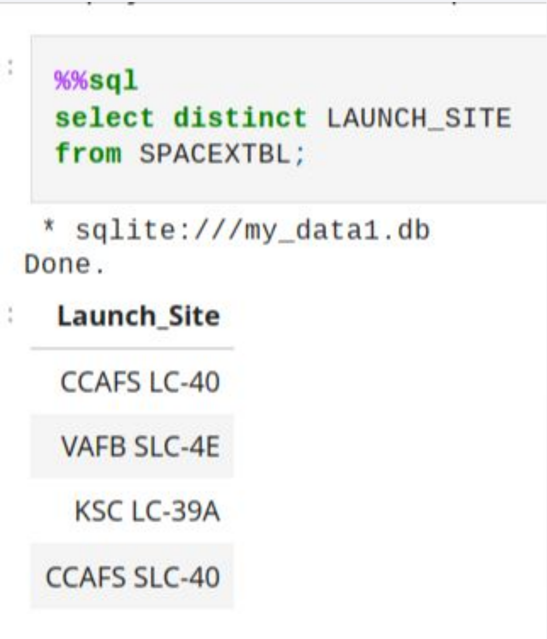
Launch Success Yearly Trend

- ➔ The success rate increased between 2013 and 2017, as well as between 2018 and 2019.
- ➔ The success rate declined from 2017 to 2018 and from 2019 to 2020.



All Launch Site Names

- ➔ An SQL table named SPACEXTBL was created from the existing DataFrame.
- ➔ DISTINCT keyword was used to identify unique launch sites.



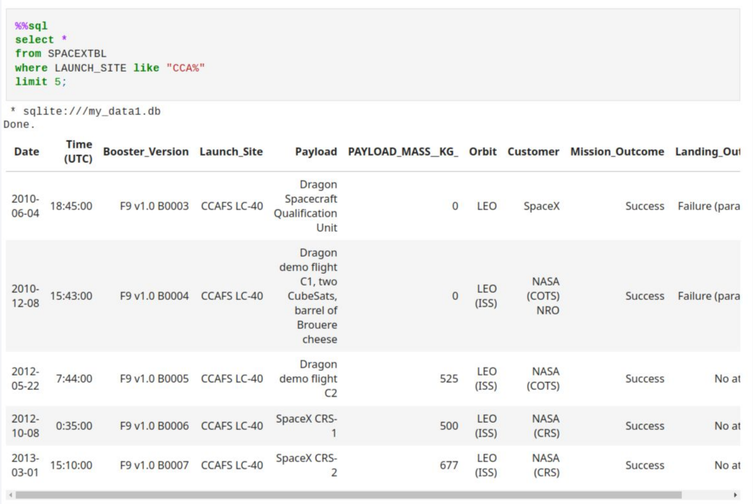
```
%%sql
select distinct LAUNCH_SITE
from SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

- ➔ The keyword LIKE CCA% was used to retrieve launch site names starting with "CCA."
- ➔ LIMIT 5 was applied to display only the first five records.



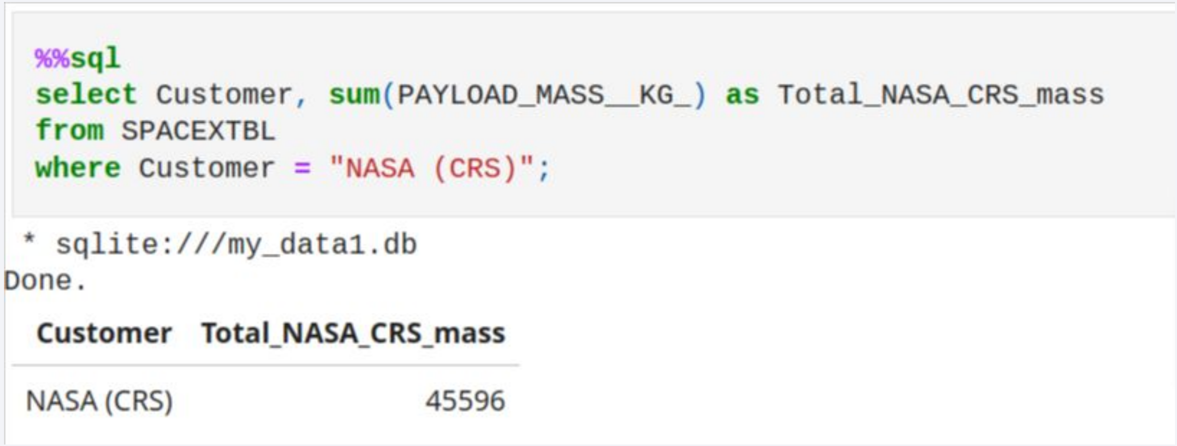
```
%%sql
select *
from SPACEXTBL
where LAUNCH_SITE like "CCA%"
limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Out
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
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 (para
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No at
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No at
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No at

Total Payload Mass

➔ The SUM function was applied to determine the total payload mass for 'NASA (CRS)' customers.



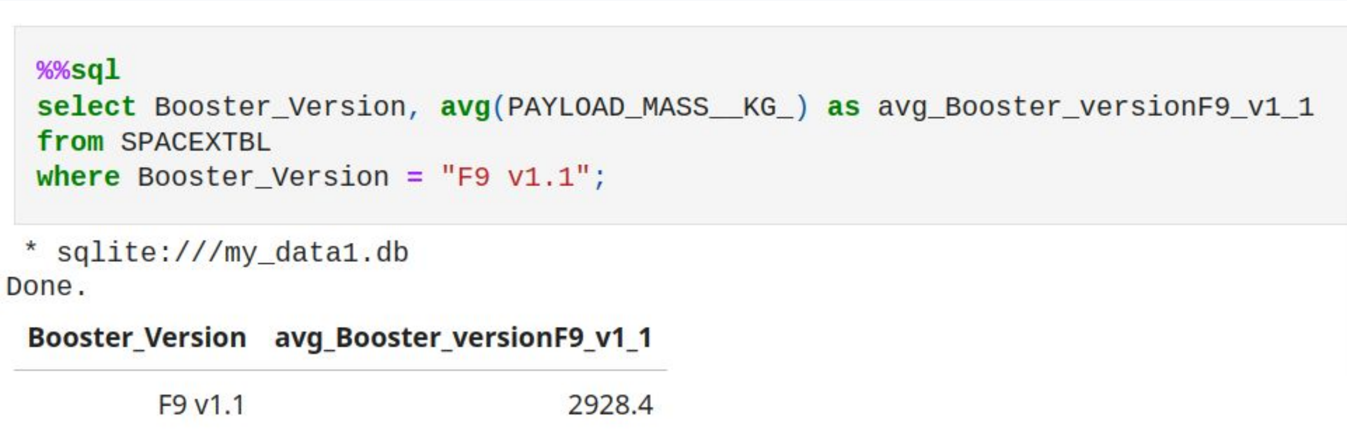
```
%%sql
select Customer, sum(PAYLOAD_MASS__KG_) as Total_NASA_CRS_mass
from SPACEXTBL
where Customer = "NASA (CRS)";
```

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

Customer	Total_NASA_CRS_mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

➔ The AVG function was applied to determine the average payload mass for booster version F9 v1.1.



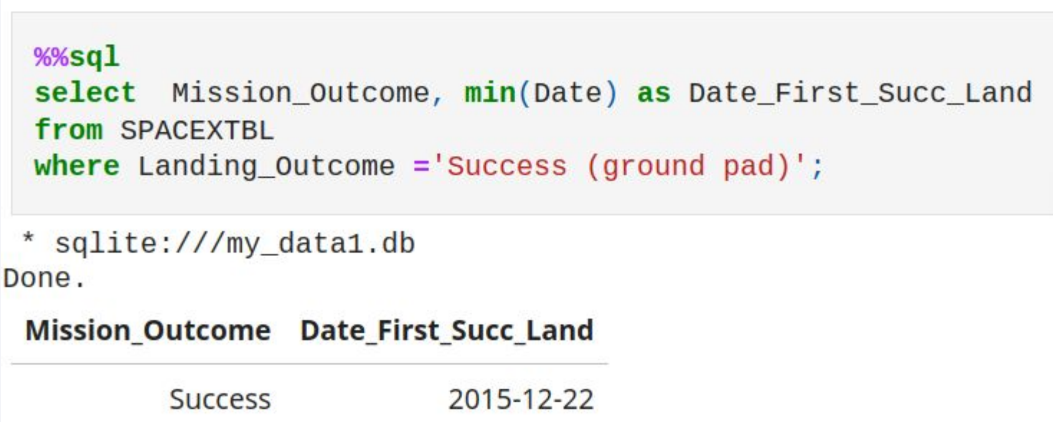
```
%%sql
select Booster_Version, avg(PAYLOAD_MASS__KG_) as avg_Booster_versionF9_v1_1
from SPACEXTBL
where Booster_Version = "F9 v1.1";
```

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

Booster_Version	avg_Booster_versionF9_v1_1
F9 v1.1	2928.4

First Successful Ground Landing Date

- ➔ An SQL query was executed to find the first successful landing on a ground pad.
- ➔ The result indicated the first successful ground landing occurred on December 22, 2015.



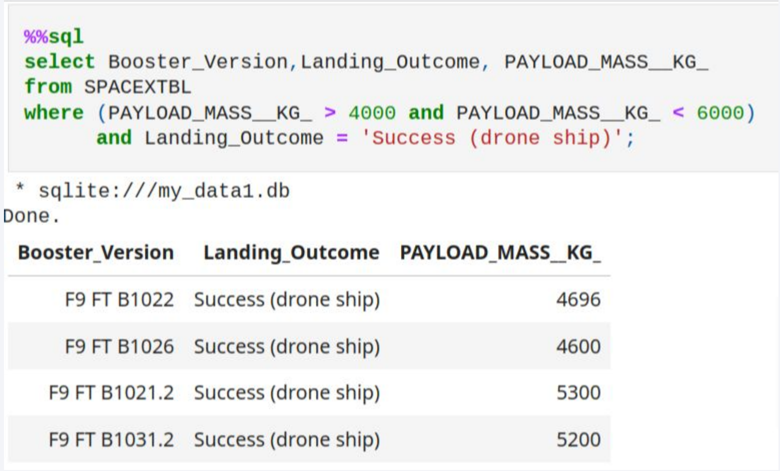
```
%%sql
select  Mission_Outcome, min(Date) as Date_First_Succ_Land
from SPACEXTBL
where Landing_Outcome ='Success (ground pad)';
```

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

Mission_Outcome	Date_First_Succ_Land
Success	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- ➔ The BETWEEN and AND keywords were used to display booster names that successfully landed on a drone ship with payload mass between 4000kg and 6000kg.
- ➔ The query returned 4 rockets.



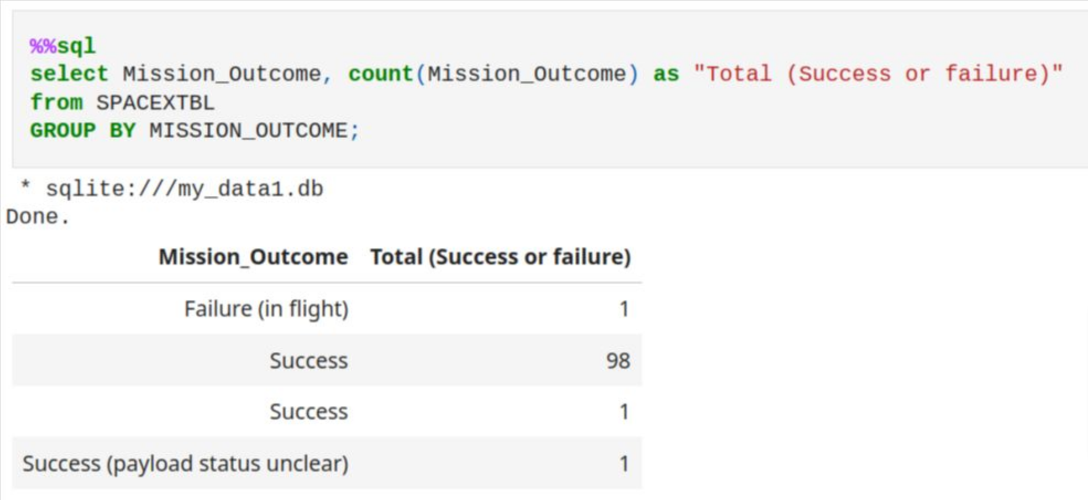
```
%%sql
select Booster_Version,Landing_Outcome, PAYLOAD_MASS__KG_
from SPACEXTBL
where (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)
      and Landing_Outcome = 'Success (drone ship)';
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS_KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- ➔ The COUNT function was used to calculate the total number of successful and failed missions.
- ➔ The results showed 100 successful missions and 1 failed mission.



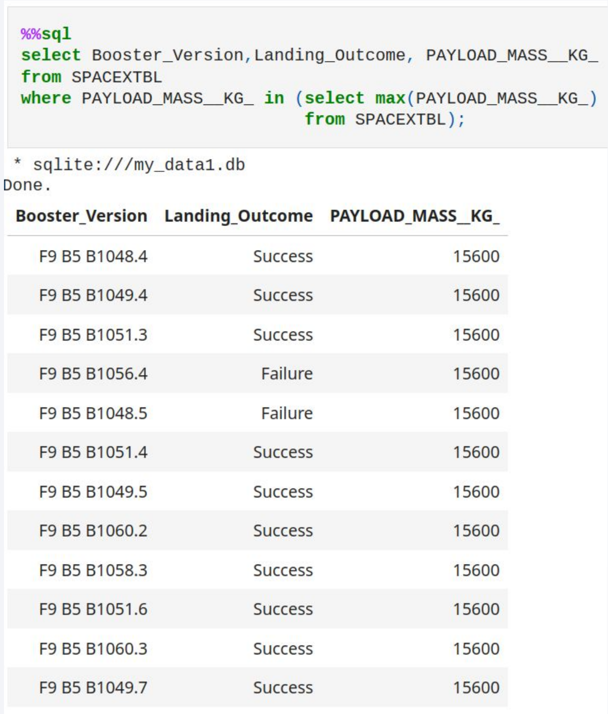
```
%%sql
select Mission_Outcome, count(Mission_Outcome) as "Total (Success or failure)"
from SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- ➔ A sub-query using the MAX function was executed to find boosters with the maximum payload.
- ➔ The results indicated a total of 12 boosters.



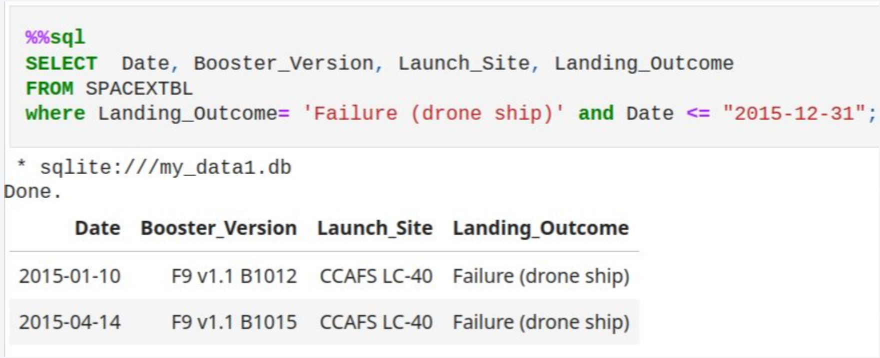
```
%%sql
select Booster_Version,Landing_Outcome, PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_)
                            from SPACEXTBL);
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Success	15600
F9 B5 B1049.4	Success	15600
F9 B5 B1051.3	Success	15600
F9 B5 B1056.4	Failure	15600
F9 B5 B1048.5	Failure	15600
F9 B5 B1051.4	Success	15600
F9 B5 B1049.5	Success	15600
F9 B5 B1060.2	Success	15600
F9 B5 B1058.3	Success	15600
F9 B5 B1051.6	Success	15600
F9 B5 B1060.3	Success	15600
F9 B5 B1049.7	Success	15600

2015 Launch Records

- ➔ The SUBSTR() function extracted the month and year from the Date column.
- ➔ The WHERE and AND keywords were applied to filter for failed drone ship landings in 2015.
- ➔ The results indicated that the failures occurred in April (04) and October (10).



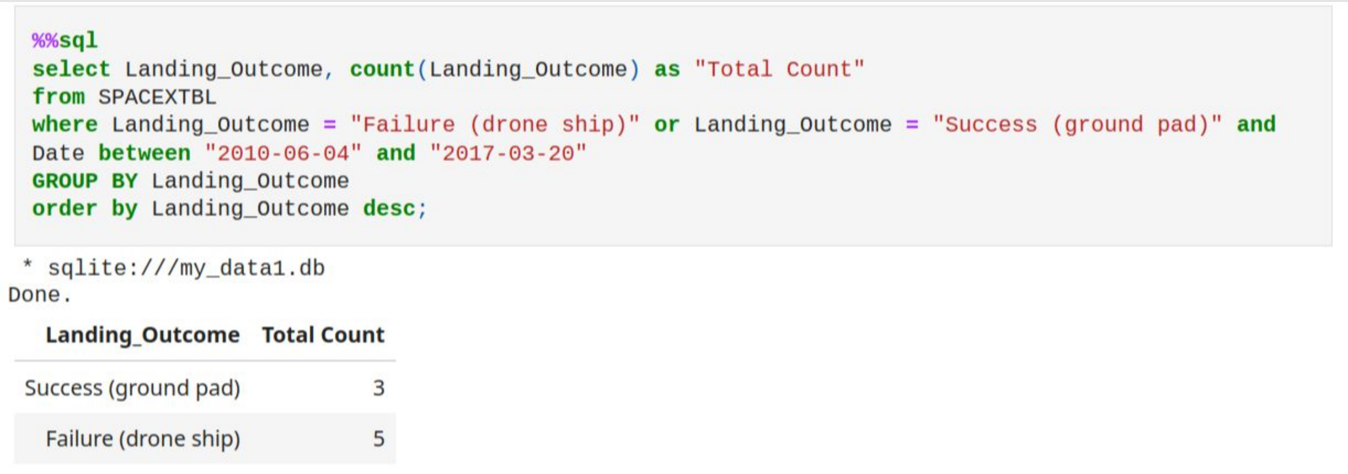
```
%%sql
SELECT  Date, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTBL
where Landing_Outcome= 'Failure (drone ship)' and Date <= "2015-12-31";
```

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

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

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

- Rank the counts of landing outcomes (e.g., Failure on drone ship or Success on ground pad) between June 4, 2010, and March 20, 2017, in descending order.



```
%%sql
select Landing_Outcome, count(Landing_Outcome) as "Total Count"
from SPACEXTBL
where Landing_Outcome = "Failure (drone ship)" or Landing_Outcome = "Success (ground pad)" and
Date between "2010-06-04" and "2017-03-20"
GROUP BY Landing_Outcome
order by Landing_Outcome desc;
```

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

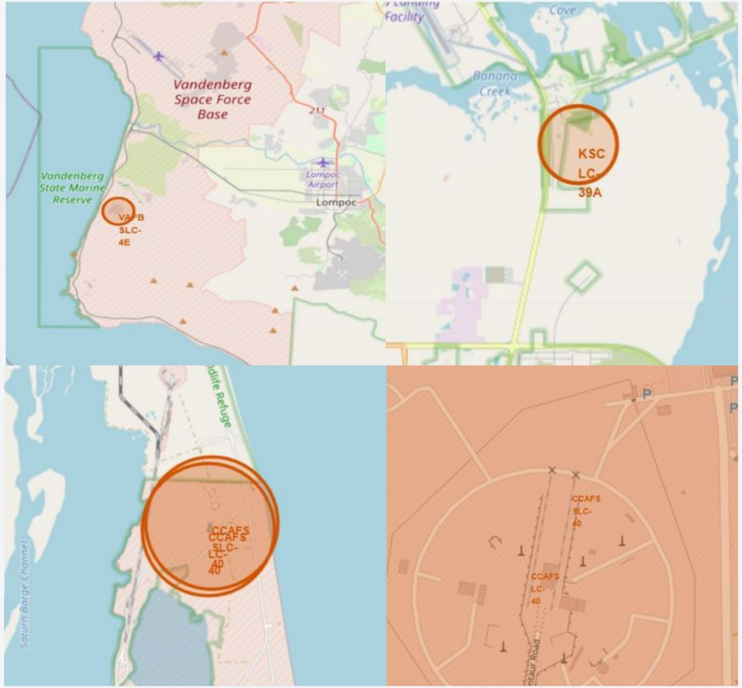
Landing_Outcome	Total Count
Success (ground pad)	3
Failure (drone ship)	5

Section
3

Launch Sites Proximities Analysis

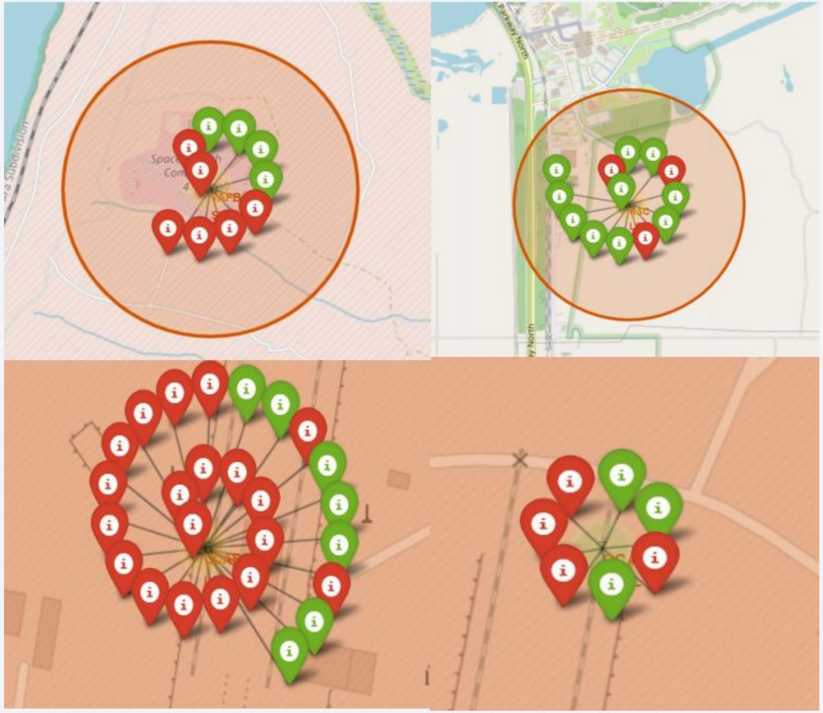
Locations of Launch Sites for SpaceX Falcon 9

➔ All launch sites depicted in the figure are situated in coastal cities across the United States.



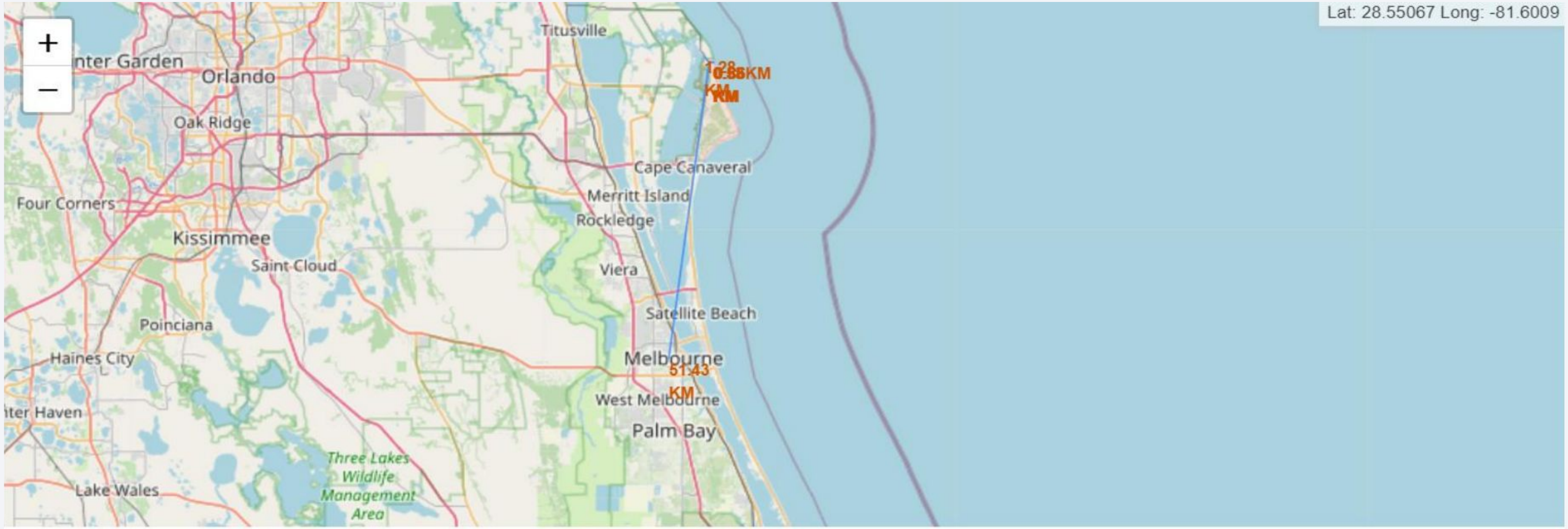
Launch Outcomes for SpaceX Falcon 9

- ➔ The figure displays the launch outcomes for various sites:
  - Top left: VAFB SLC-4E
  - Top right: KSC LC-39A
  - Bottom left: CCAFS SLC-40
  - Bottom right: CCAFS LC-40
- ➔ Red icons indicate failures, while green icons signify successes.



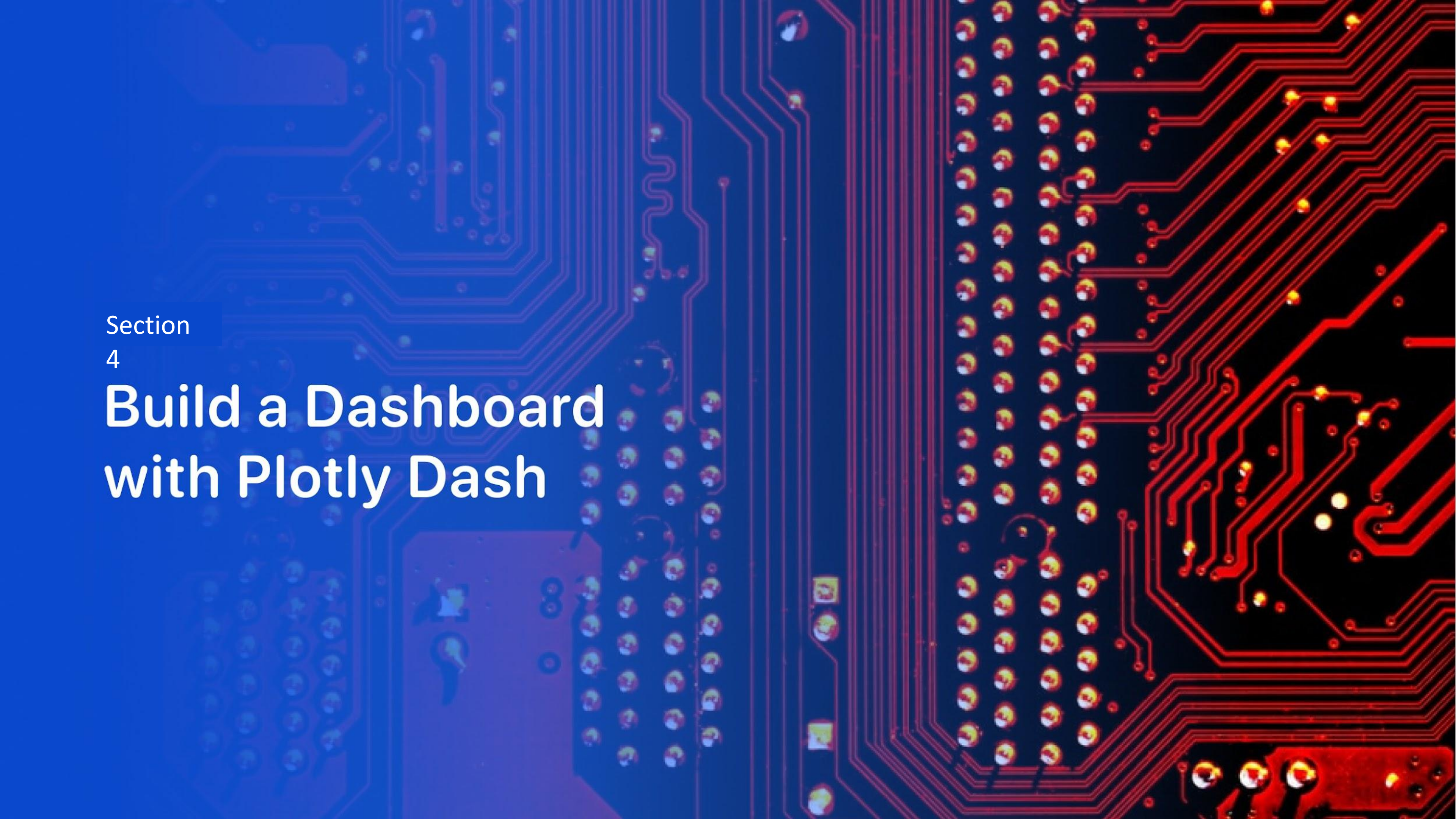
Proximities

- ➔ Proximities, including railways, highways, and coastlines, with calculated distances shown.



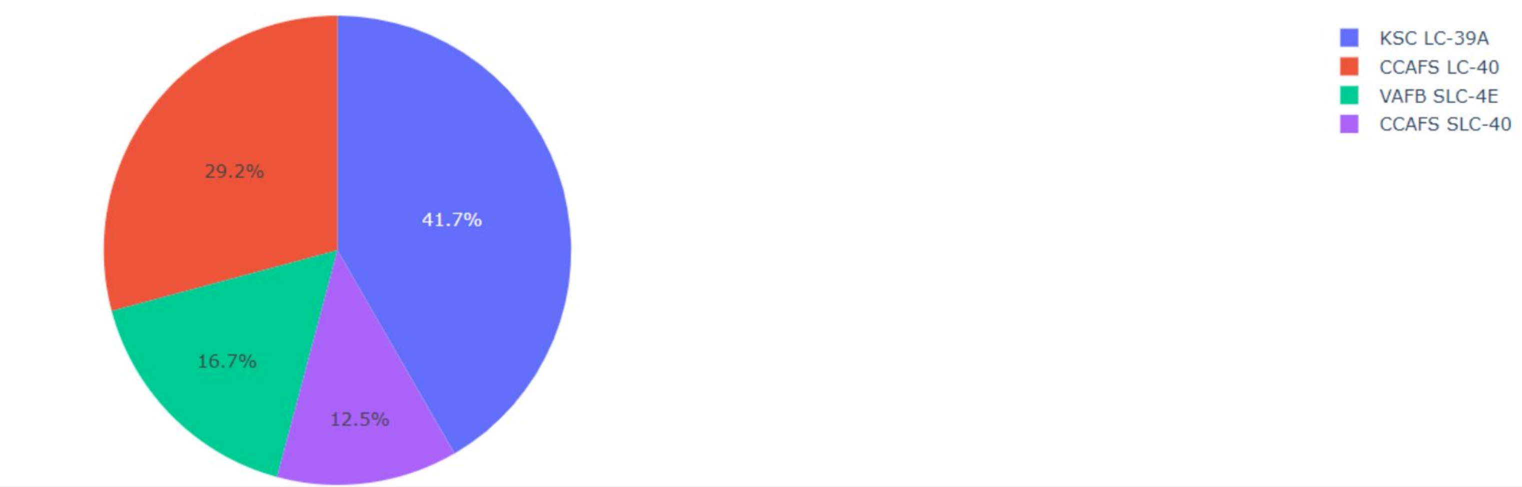
Section
4

Build a Dashboard with Plotly Dash



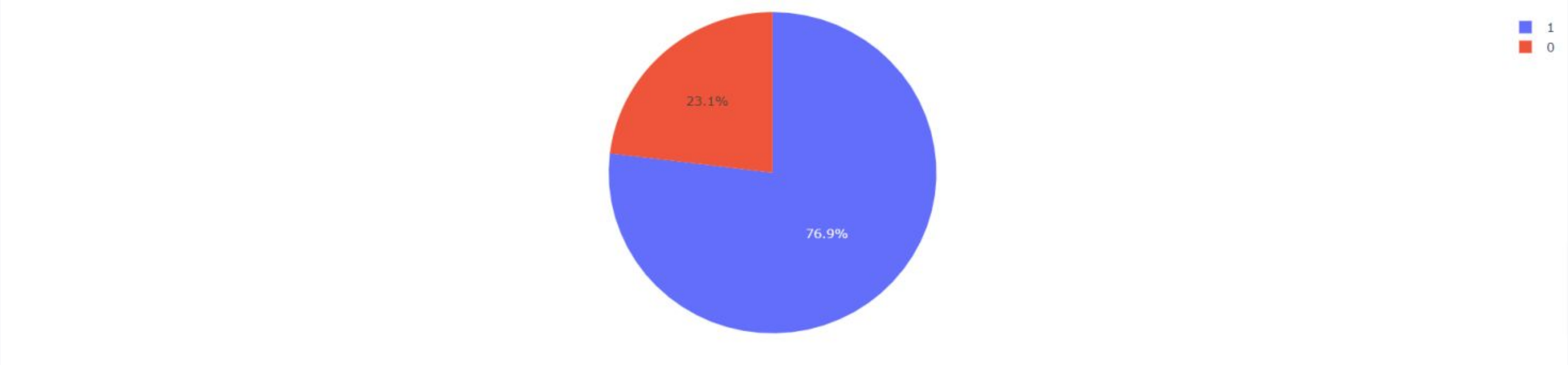
Pie chart representing all launch sites

- ➔ KSC LC-39A accounts for 41.7% of the pie chart, indicating it has more launches than the other sites, followed by CCAFS LC-40 at 29.2%.



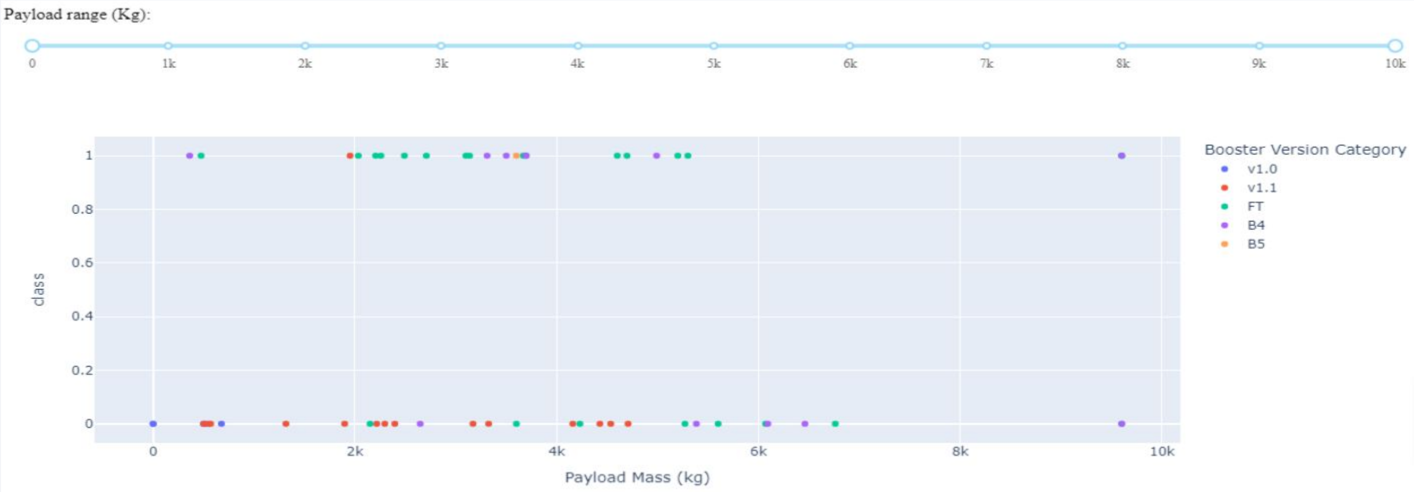
Pie chart of KSC LC 39A

➔ KSC LC-39A has the highest count, boasting a 76.9% success rate for Falcon 9 landings.



Payload vs. Launch Outcomes for All Sites

➔ The scatter plot reveals that as payload mass increases, the success rate for the booster version FT also rises, whereas the opposite trend is observed for the booster version v1.1, which experiences more failures.

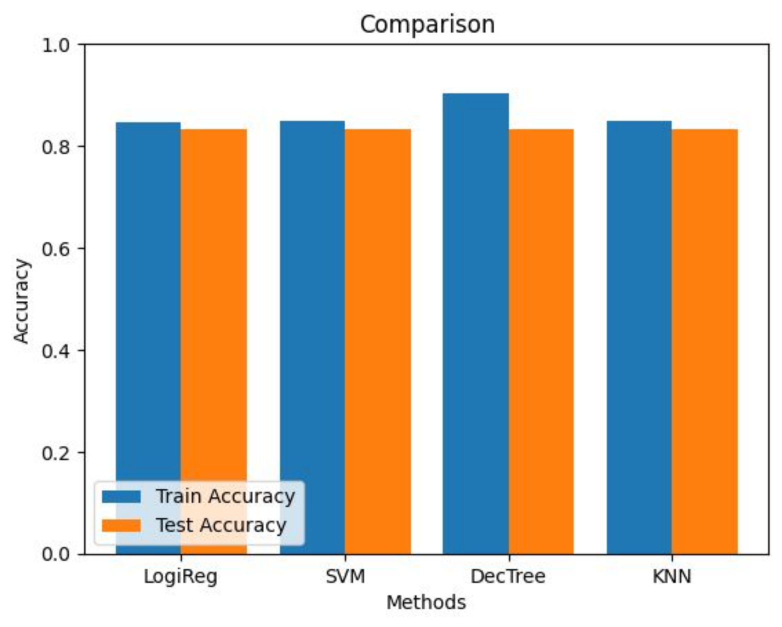


Section
5

Predictive Analysis (Classification)

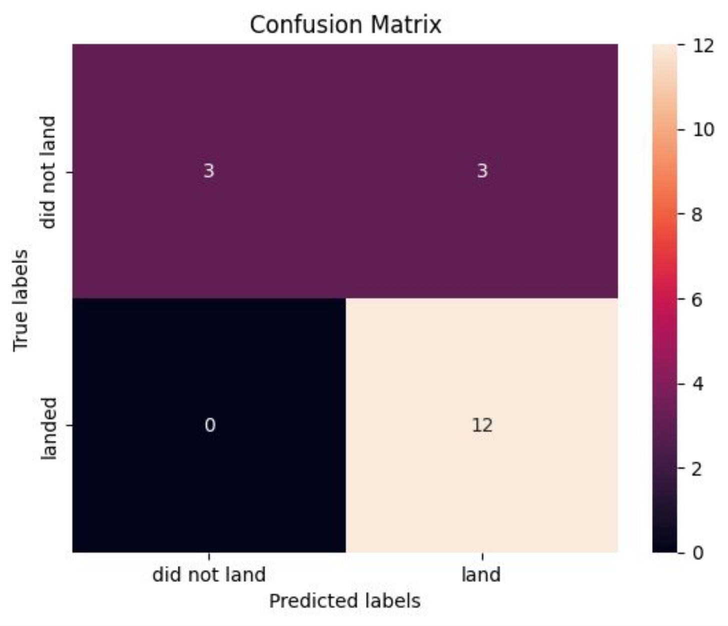
Classification Accuracy

From the bar chart, Decision tree classifier performed slightly better.



Confusion Matrix

- ➔ After splitting the dataset into training and test sets, only 18 test samples remained.
- ➔ The decision tree classifier accurately predicted 12 successful landings (12 true positives) and 3 unsuccessful landings (3 true negatives) from the test set.
- ➔ There were no false negatives, as the classifier did not misclassify any successful landings.
- ➔ However, it produced 3 false positives, incorrectly predicting that 3 observations were successful.



Conclusions

- ➔ An increase in flight numbers correlates with a higher success rate for Launch Site CCAFS SLC-40 and similarly for payload mass.
- ➔ Orbit types ES-L1, SSO, HEO, and GEO achieve a 100% landing success rate.
- ➔ The success rate of Falcon 9 landings has shown a steady increase from 2010 to 2020.
- ➔ Successful missions tend to have smaller payload masses, which should be factored into predictions.
- ➔ Decision tree classifiers outperformed other classification algorithms, achieving approximately 87% accuracy in predicting landing outcomes.

Appendix

- All relevant assets can be found on Github.

Thank you!