Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary: Methodologies Overview

The research is focused on identifying key factors that contribute to successful rocket landings. The following steps were undertaken to achieve this goal:

1.Data Collection
- ◦Utilized SpaceX REST API and web scraping tools to gather relevant data.

2.Data Wrangling
- ◦Processed and structured data to create a success/failure outcome variable.

3.Data Exploration
- ◦Employed data visualization techniques to analyze variables such as:
  - •Payload, launch site, flight number, and annual trends.

4.Data Analysis
- ◦Used SQL to compute key statistics, including:
  - •Total payload, payload ranges for successful launches, and counts of successful and failed outcomes.

5.Exploration of Launch Sites
- ◦Examined success rates and their proximity to geographical markers.

6.Data Visualization
- ◦Highlighted launch sites with the highest success rates and optimal payload ranges.

7.Model Building
- ◦Developed predictive models (logistic regression, support vector machines (SVM), decision trees, and K-nearest neighbor (KNN)) to forecast landing outcomes.

Executive Summary: Key Results

1. Exploratory Data Analysis

- Rocket launch success has improved over time.
- KSC LC-39A is the site with the highest success rate.
- Specific orbits (e.g., ES-L1, GEO, HEO, SSO) consistently show a 100% success rate.

2. Visualization & Analytics

- Most launch sites are situated near the equator and close to coastal areas.

3. Predictive Analytics

- All models performed comparably on test data, with the decision tree model showing slightly better results.

Introduction

Background:
SpaceX aims to make space travel affordable, achieving milestones like ISS missions, satellite constellations for global internet, and manned spaceflights. Its reusable Falcon 9 first stage reduces launch costs to $62M, compared to $165M for non-reusable competitors.

Objective:
Use public data and machine learning to predict first-stage landing success, which directly impacts launch costs.
Key Focus Areas:
1.Analyze how payload mass, launch site, flights, and orbit type affect landing success.
2.Track success trends over time.
3.Determine the best predictive model for landing success.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

Methodology: Steps to Analyze and Predict Landing Outcomes

1.Data Collection:
Use the SpaceX REST API and web scraping to gather relevant data.
1.Data Wrangling:
- ◦Filter and clean data.
- ◦Handle missing values.
- ◦Apply one-hot encoding to prepare for analysis and modeling.

2.Exploratory Data Analysis (EDA):

Perform EDA using SQL queries and data visualization techniques.

1.Data Visualization:
Create visualizations with Folium and Plotly Dash to better understand patterns.
1.Model Development:
- ◦Build classification models to predict landing outcomes.
- ◦Tune and evaluate models to identify the best-performing model and parameters.

Data Collection – SpaceX API

Request Data:
Fetch rocket launch data from the SpaceX API.

Process Response:
•Decode the API response using .json() && Convert the response into a DataFrame with .json_normalize().

Retrieve Launch Details:
Use custom functions to request detailed launch information.

Organize Data:
•Create a dictionary from the collected data.
•Convert the dictionary into a DataFrame.

Filter Data:
Retain only Falcon 9 launch records.

Handle Missing Values:
Replace missing values in the Payload Mass column with the calculated mean.

Export Data:
Save the cleaned dataset as a CSV file for further analysis.

Data Collection - Scraping

Request Data:
Fetch Falcon 9 launch data from Wikipedia.

Parse HTML Response:
Create a BeautifulSoup object from the HTML content.

Extract Table Headers:
Identify and extract column names from the HTML table header.

Parse Table Data:
Collect data by parsing the rows and cells of the HTML tables.

Organize Data:
•Create a dictionary from the parsed data.
•Convert the dictionary into a DataFrame.

Export Data:
Save the DataFrame to a CSV file for further analysis.

Data Wrangling

Perform EDA:
Analyze the data and determine labels for the dependent variable (landing outcome).
Calculate Metrics:
- ◦Total launches for each site.
- ◦Count and frequency of orbit types.
- ◦Count and frequency of mission outcomes per orbit type.

Create Binary Landing Outcome:
- ◦Define the landing outcome column:
  - •True Ocean: Successful landing in a specific ocean region.
  - •False Ocean: Unsuccessful landing in the ocean.
  - •True RTLS: Successful landing on a ground pad.
  - •False RTLS: Unsuccessful landing on a ground pad.
  - •True ASDS: Successful landing on a drone ship.
  - •False ASDS: Unsuccessful landing on a drone ship.
- ◦Convert outcomes to binary: 1 for success, 0 for failure.

Export Data:
Save the processed dataset as a CSV file for further analysis.
Notes:
•Landing success was not guaranteed for all missions.
•True Ocean represented a successful landing in an ocean region.

EDA with Data Visualization

Charts to Create

1.Flight Number vs. Payload:
Use a scatter plot to observe how payload mass varies with the flight number.

2. Flight Number vs. Launch Site:
Plot a scatter plot to examine how flight numbers are distributed across launch sites.

3. Payload Mass (kg) vs. Launch Site:
Create a bar chart to compare payload capacities across different launch sites.

4.Payload Mass (kg) vs. Orbit Type:
Use a bar chart to show how payload mass varies across different orbit types.

Analysis

•**Scatter Plots:**
Reveal relationships between numerical variables. For example:

- ◦Flight number trends may indicate improvements in payload capacity over time.
- ◦Comparing payload mass across launch sites might show site-specific payload capabilities.

•**Bar Charts:**
Useful for categorical comparisons, such as:

- ◦Launch site performance and payload mass distributions.
- ◦Orbit type preferences for heavier or lighter payloads.

Objective: Identify patterns or relationships that could be relevant for machine learning models.

EDA with SQL

Date of First Successful Landing on Ground Pad

Identify the date when SpaceX achieved its first successful ground pad landing.

Boosters with Successful Drone Ship Landings (Payload Mass Between 4,000 and 6,000)

List boosters that successfully landed on a drone ship with payload mass between 4,000 kg and 6,000 kg.

Total Number of Successful and Failed Missions

Count the total number of successful and failed SpaceX missions.

Boosters with Max Payload Carried

Identify which booster versions carried the maximum payload.

Failed Drone Ship Landings in 2015 (Booster Version and Launch Site)

List the failed landing outcomes on a drone ship in 2015, including booster versions and launch sites.

Count of Landing Outcomes Between 2010-06-04 and 2017-03-20 (Descending Order)

Count and display the landing outcomes between June 4, 2010, and March 20, 2017, sorted in descending order of frequency.

https://github.com/davidlanguage/spacex_ibm/blob/main/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

1. Markers: they are specific spots on the map.
   Markers highlight important locations, like landmarks or event sites. You can also attach extra details in popups, providing more context about each place.

2. Circles: They represent an area around a central point, defined by a radius.
   Circles can show the impact of an event, like the coverage of a service, the range of a sensor, or the affected area in a disaster.

3. Lines: They connect different points on the map.
   Lines help visualize paths, routes, or connections between locations, such as travel routes, migration patterns, or even network connections.

Adding these elements to a folium map lets you create a clearer and more engaging way to present and analyze spatial data, making it easier to spot patterns and understand geographical relationships.

https://github.com/davidlanguage/spacex_ibm/blob/main/6_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Pie Chart for Launch Success by Site (get_pie_chart)
Input: `site-dropdown` (dropdown menu).
Output: `success-pie-chart` (pie chart).
Displays the total number of successful launches for each site. If "ALL" is selected, it shows the distribution across all sites. For a specific site, it shows the success rate for that site. Provides a quick visual summary of launch success rates by site.

Scatter Plot for Payload vs. Launch Success (update_scatter_chart)
Input: `site-dropdown` (dropdown menu), `payload-slider` (slider for payload mass range).
Output: `success-payload-scatter-chart` (scatter plot).
Shows the correlation between payload mass and launch success. Users can filter by site and payload range. Data points are color-coded by booster version. Helps analyze the relationship between payload mass and launch success, highlighting trends and patterns.

https://github.com/davidlanguage/spacex_ibm/blob/main/7_spacex_dash_app.py

Predictive Analysis (Classification)

Data Preparatiom:
Loaded Data: Imported SpaceX launch data.
Cleaned Data: Handled missing values and encoded categorical variables.

Model Building
Selected Algorithms
Chose Decision Tree, K-Nearest Neighbors, Logistic Regression, and Support Vector Machine.

Model Evaluation
Cross-Validation: Used GridSearchCV with `cv=10` to evaluate models.
Focused on accuracy to compare models.

Model Improvement**
- Hyperparameter Tuning**: Adjusted parameters to improve performance.
- Feature Engineering**: Added/removed features based on importance.

Best Model Selection
Compared models using cross-validation scores.
Chose the model with the highest accuracy.

Flowchart

Data Preparation
↓
Model Building
↓
Model Evaluation
↓
Model Improvement
↓
Best Model Selection

https://github.com/davidlanguage/spacex_ibm/blob/main/8_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

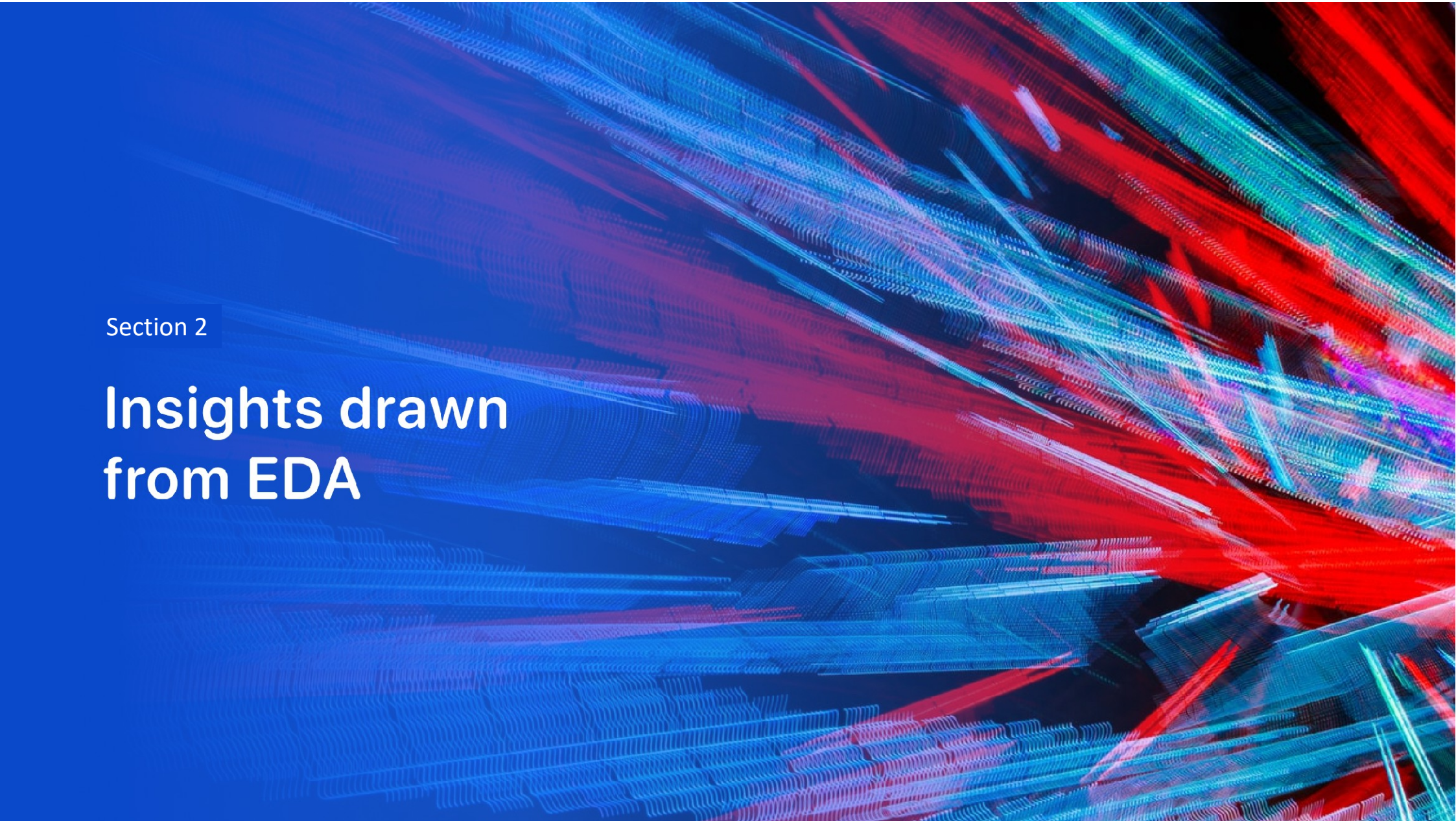
Visual Analytics

Launch Site Distribution: Most launch sites are positioned near the equator, and all are strategically located close to the coast.

Safety & Accessibility: Launch sites are far from populated areas (e.g., cities, highways, railways) to minimize damage in case of a failure. They are also conveniently located to facilitate transportation of people and materials needed for launch activities.

Predictive Analytics

Best Model: The Decision Tree model provides the most accurate predictions for launch success and landing outcomes in the dataset.

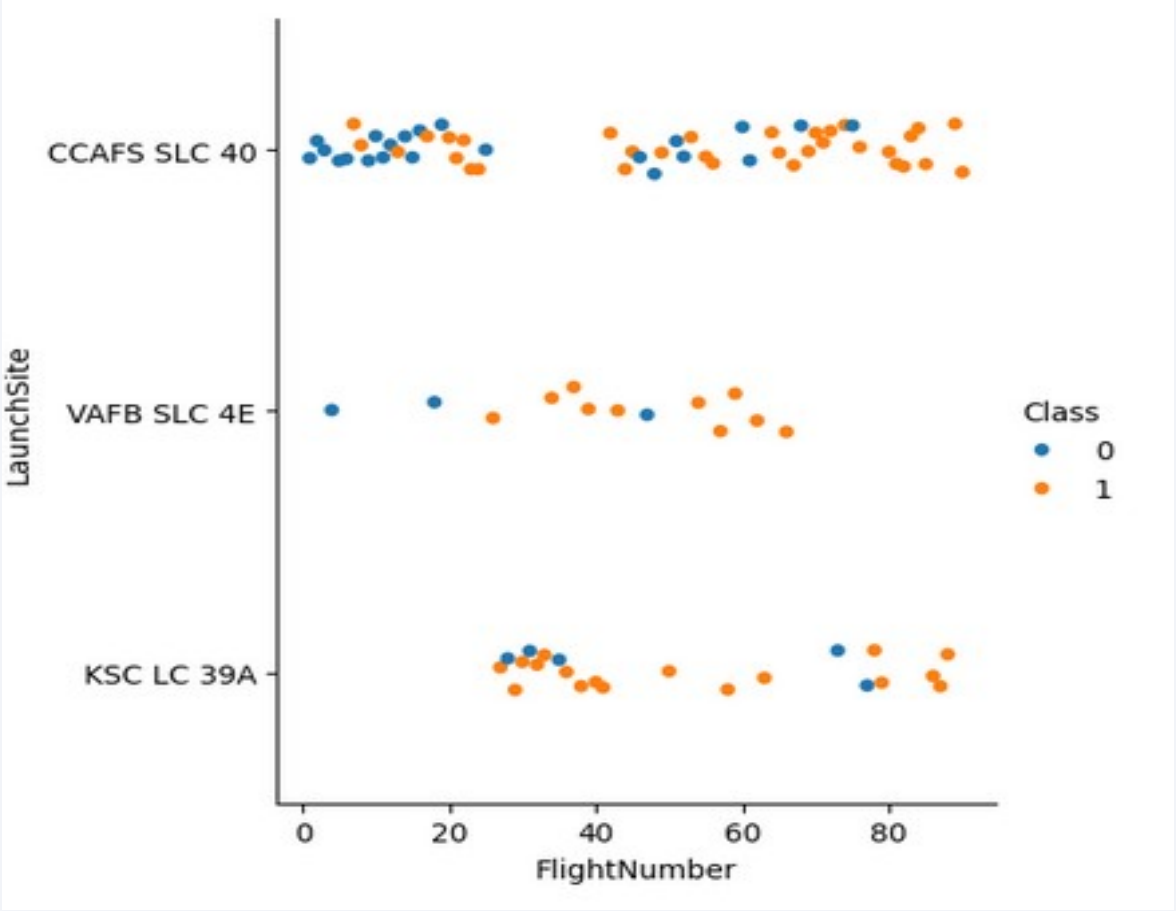


Section 2

Insights drawn from EDA

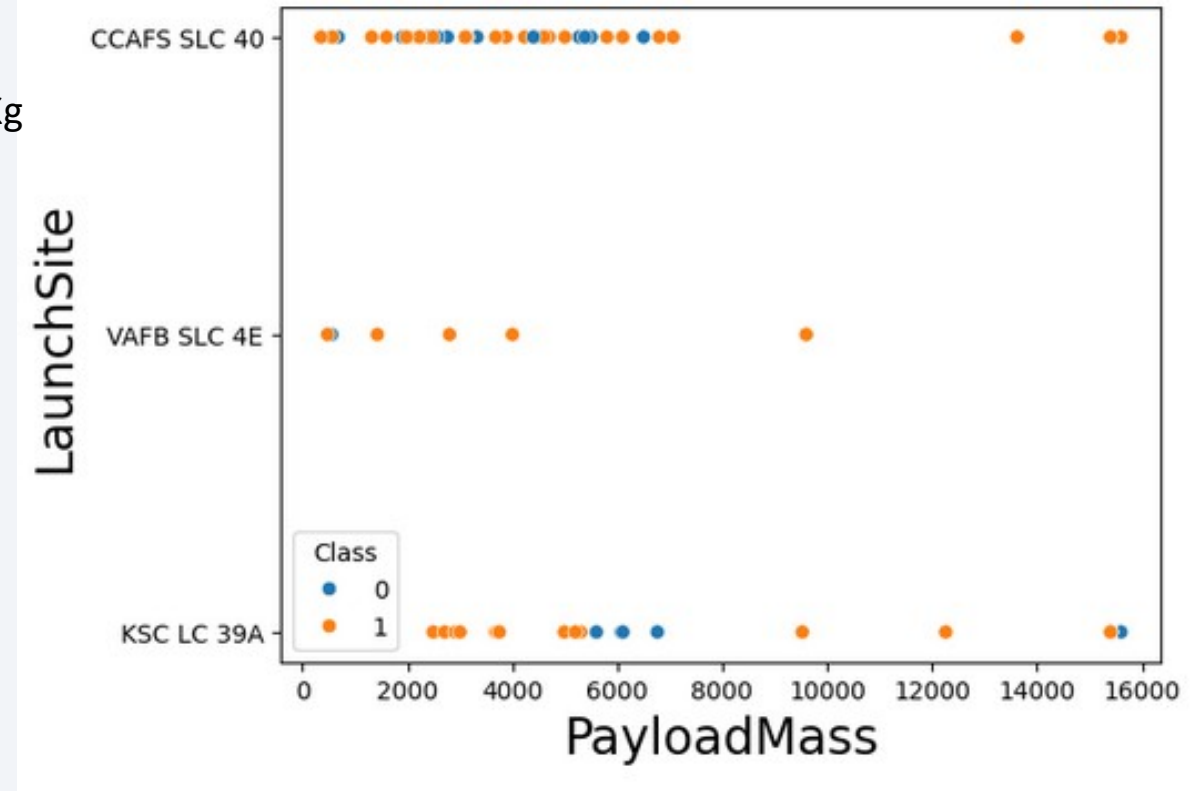
Flight Number vs. Launch Site

Class 1 (orange) = success
Class 2 (blue) = failure



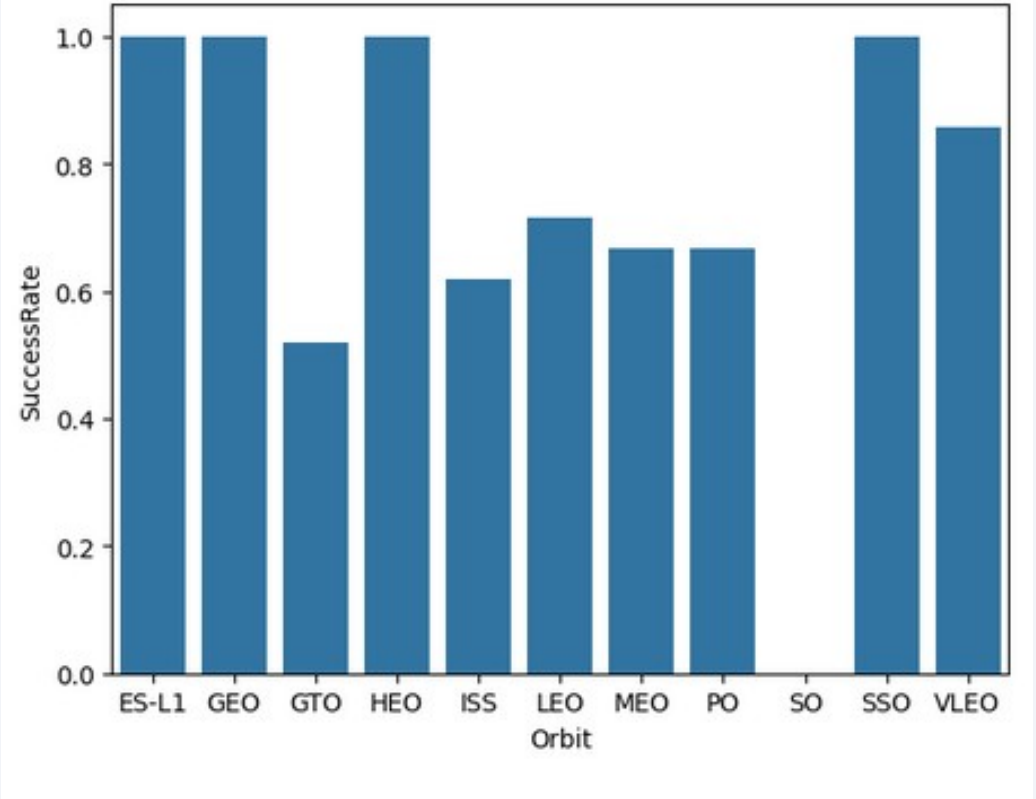
Payload vs. Launch Site

PayloadMass value is displayed in Kg

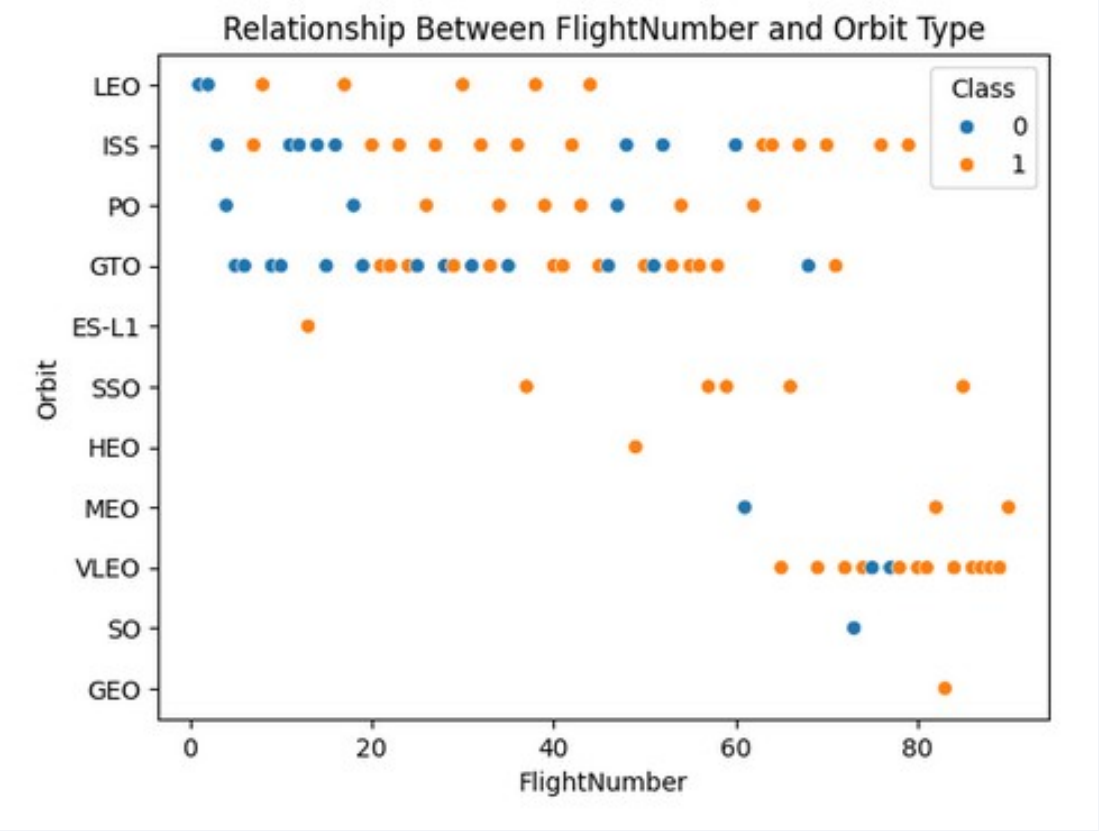


Success Rate vs. Orbit Type

Value range 1 to 0 represent a success rate from 100% to 0%

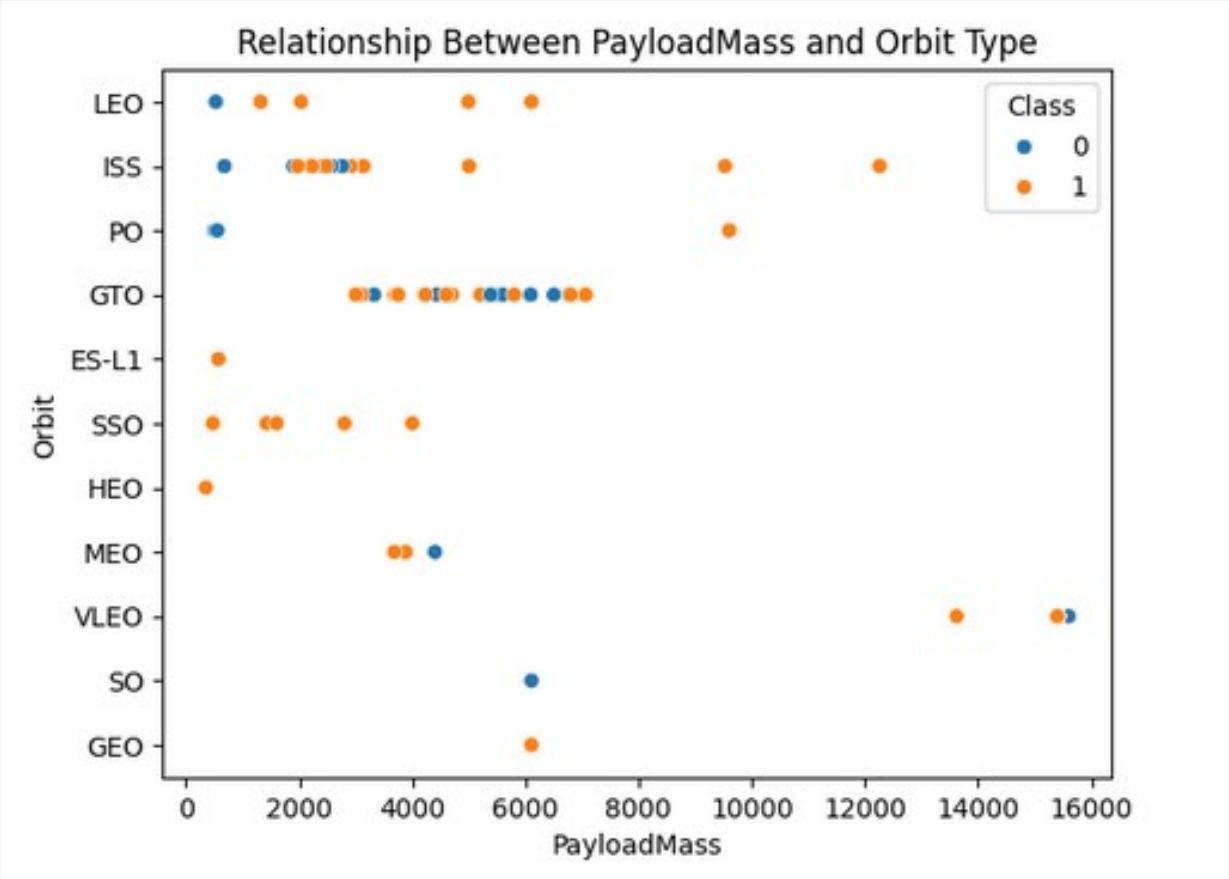


Flight Number vs. Orbit Type



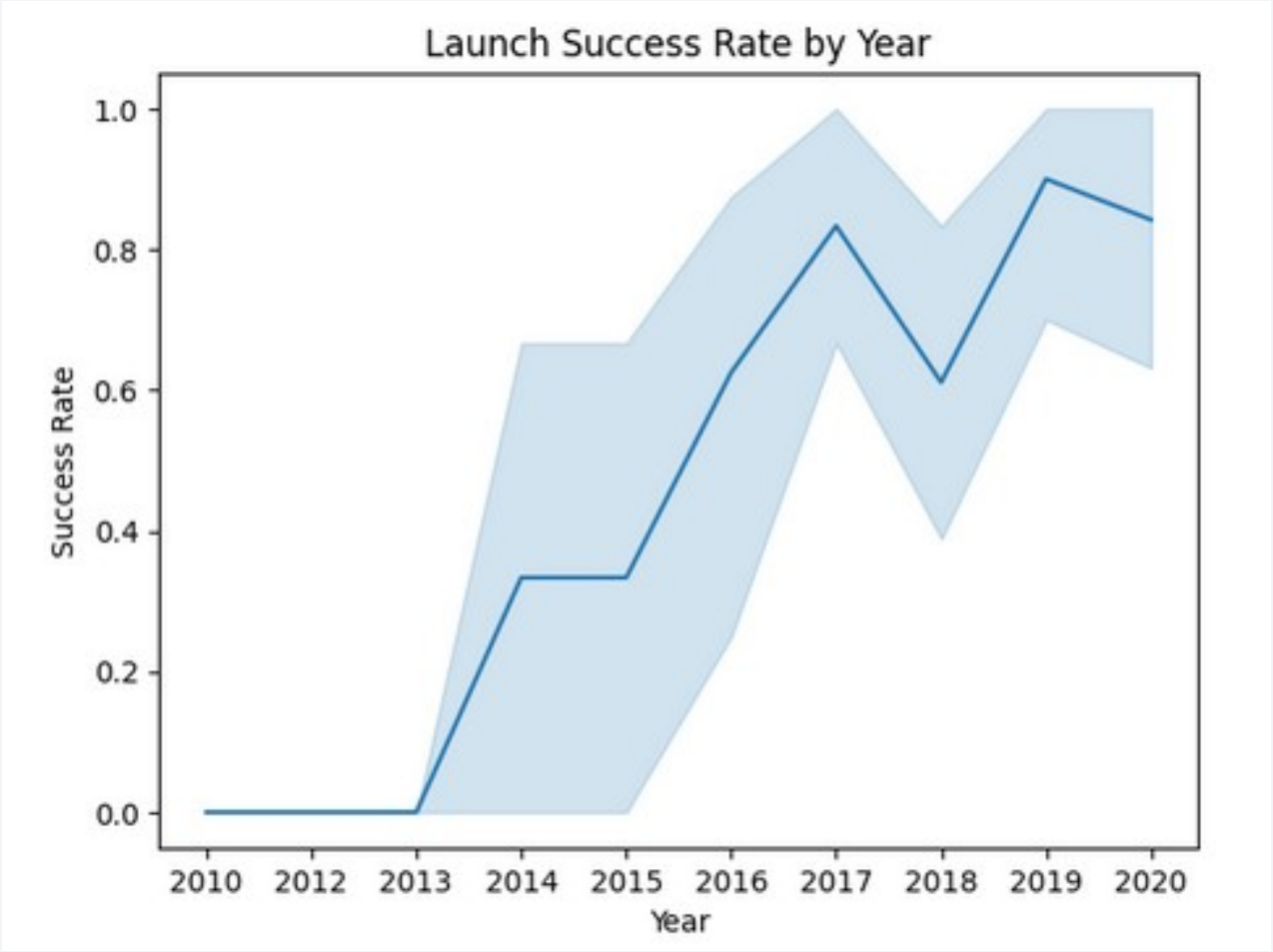
Payload vs. Orbit Type

PayloadMass value is displayed in Kg



Launch Success Yearly Trend

Success value from 1 to 0, represent from 100% to 0% success



All Launch Site Names

Launch Site Unique Names
1. CCAFS LC-40
2. CCAFS SLC-40
3. KSC LC-39A
4. VAFB SLC-4E

Query Result

	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...
0	1	6104.959412	1	False	False	False	1.0	0	False	False	...
1	2	525.000000	1	False	False	False	1.0	0	False	False	...
2	3	677.000000	1	False	False	False	1.0	0	False	False	...
3	4	500.000000	1	False	False	False	1.0	0	False	False	...
4	5	3170.000000	1	False	False	False	1.0	0	False	False	...

Launch Site Names Begin with 'CCA'

Query and Result displayed from code in the screenshot next to this text.

LIKE ‘CCA%’ contains the ‘%’ placeholder, indicating any additional char can follow the CCA value.

LIMIT 5 restrains the number of results to 5, ensuring no further entries are shown, even when they exist.

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

```
%sql SELECT * FROM SPACEXTABLE 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_C
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)'
```

[46] ✓ 0.0s Python

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Tanks to build-in AVG() we can directly get the average in a single line.

With WHERRE Booster_Version = ‘F9 v1.1’, we are exclusively getting the records with the specific type of booster.

Task 4

Display average payload mass carried by booster version F9 v1.1

%sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'

```
[12]: %sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
t[12]: AVG(PAYLOAD_MASS__KG_)
       2928.4
```

First Successful Ground Landing Date

```
In [14]: %sql SELECT Min(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
Out[14]: Min(Date)
         2015-12-22
```

MIN() ensures we get the minimum value from a series of records

WHERE Landing_Outcome =? ‘Success (ground pad)’ limits the returned results to those with a successful ground landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

0.0s Python

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

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

BETWEEN 4000 AND 6000 is key to limit the Payload Mass required in our results. With they WHERE clause, we are limiting ourselves to those with a successful drone ship.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT SUM(CASE WHEN Mission_Outcome = 'Success' THEN 1 ELSE 0 END) as Success_Count, SUM(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 ELSE 0 END) as Failure_Count FROM SPACEXTABLE;
```

✓ 0.0s

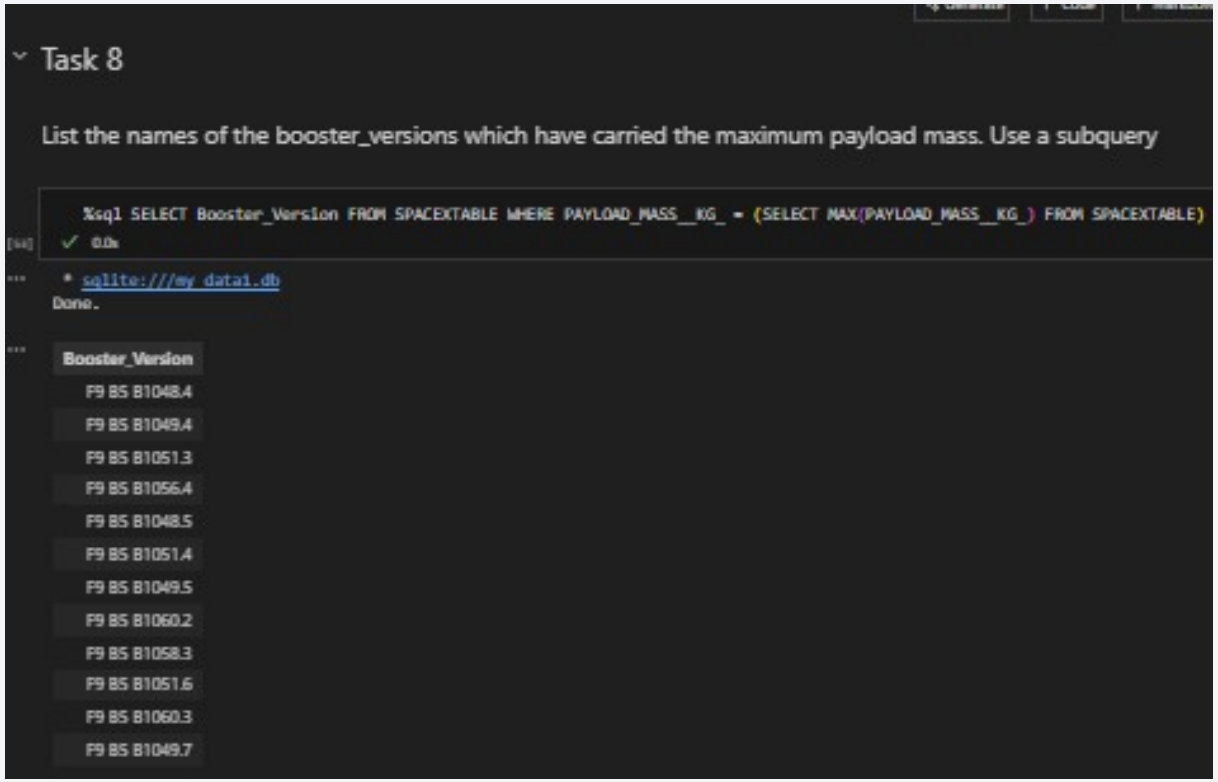
```
 * sqlite:///my_data1.db
Done.
```

Success_Count	Failure_Count
98	1

Task 9

CASE WHEN is needed to create a personalized boolean return statement.

Boosters Carried Maximum Payload



2015 Launch Records

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 substr(Date,0,5) = '2015' AND Landing_Outcome LIKE '%drone ship%'
```

✓ 0.0s

```
* 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
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

substr() ensures we get the month value. WHERE substr(Date,0,5) take care of ensure we cosider only the beginning from date value, for the year.

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

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(Landing_Outcome) as Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Count DESC
```

✓ 0.0s

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

Landing_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

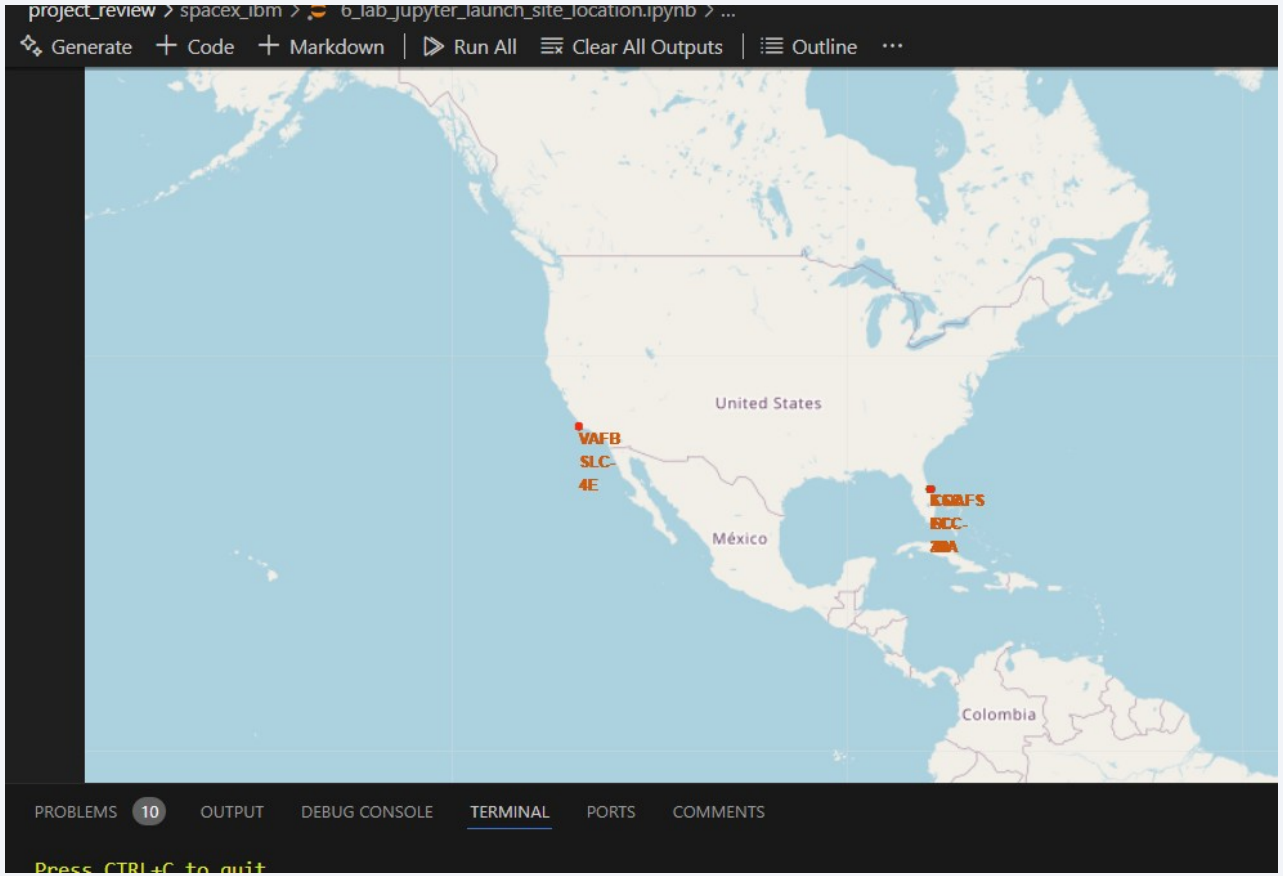
Section 3

Launch Sites Proximities Analysis

Equatorial Launch Sites

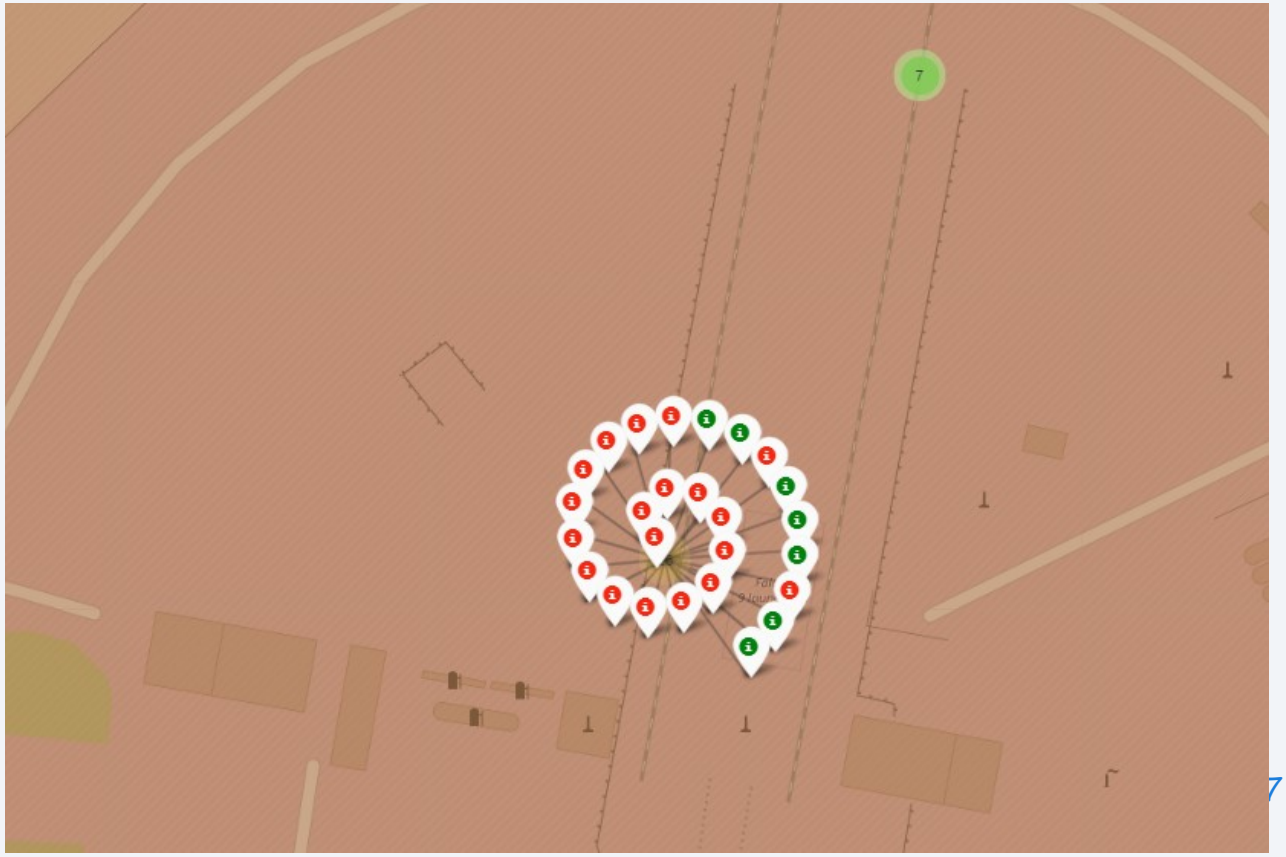
Launch Sites are close the equator. There, it is easiest it is to launch to the equatorial orbit.

Additionally, thanks to theEarth's rotation, rockets launched from sites near the equator get an additional boost, thus saving on fuel costs.



Launch Outcomes

Analyzing launching outcomes provides an useful insight in optimal launch location for future missions.



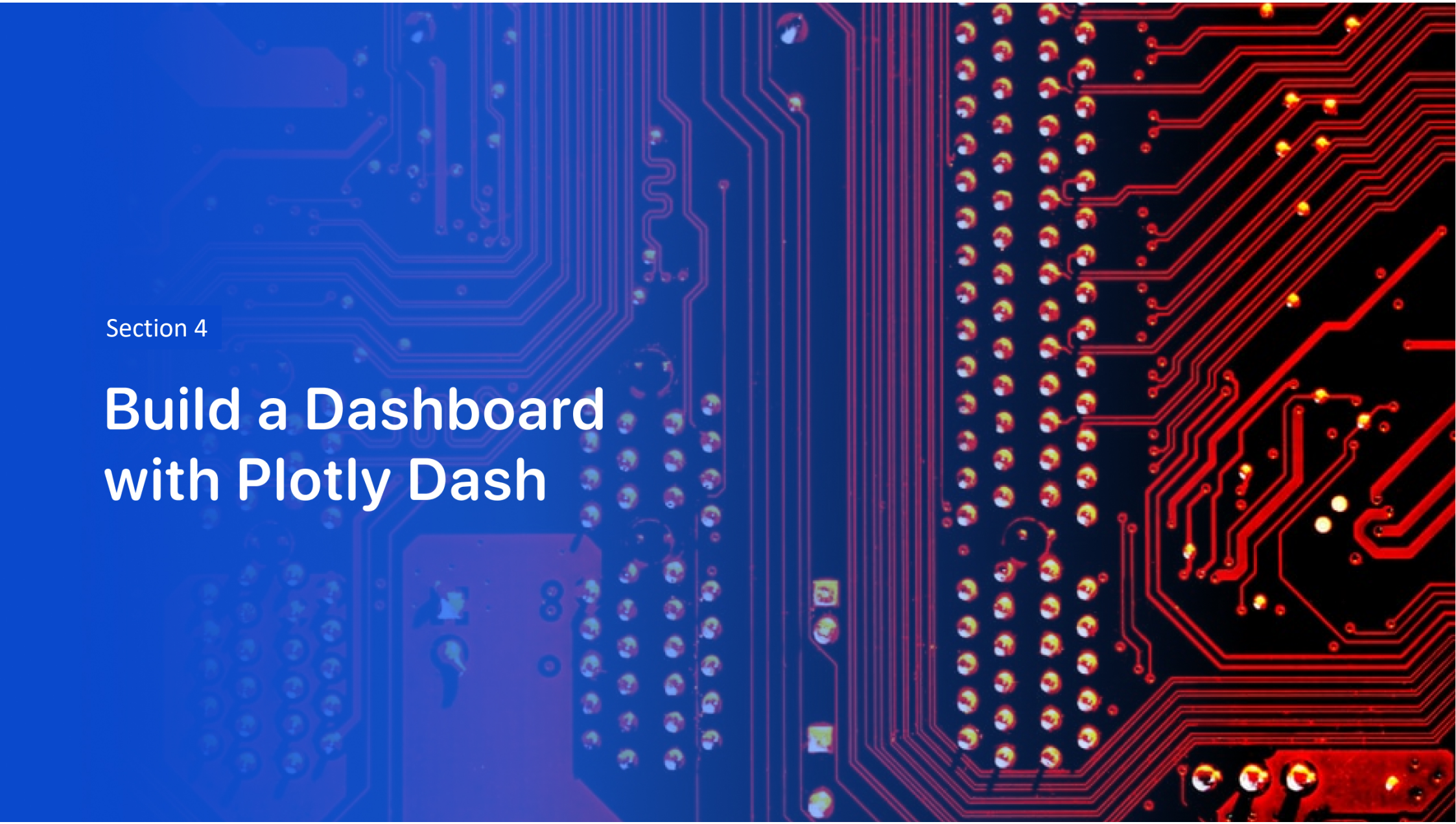
Launch Proximities

Why Launch Sites Are Near Coasts and Isolated Areas?

Ensures that spent rocket stages or debris from failed launches fall into the ocean, avoiding harm to people or property.

Launch sites require exclusion zones to prevent unauthorized access and to keep people safe during launches. These zones protect against potential hazards such as explosions or debris.

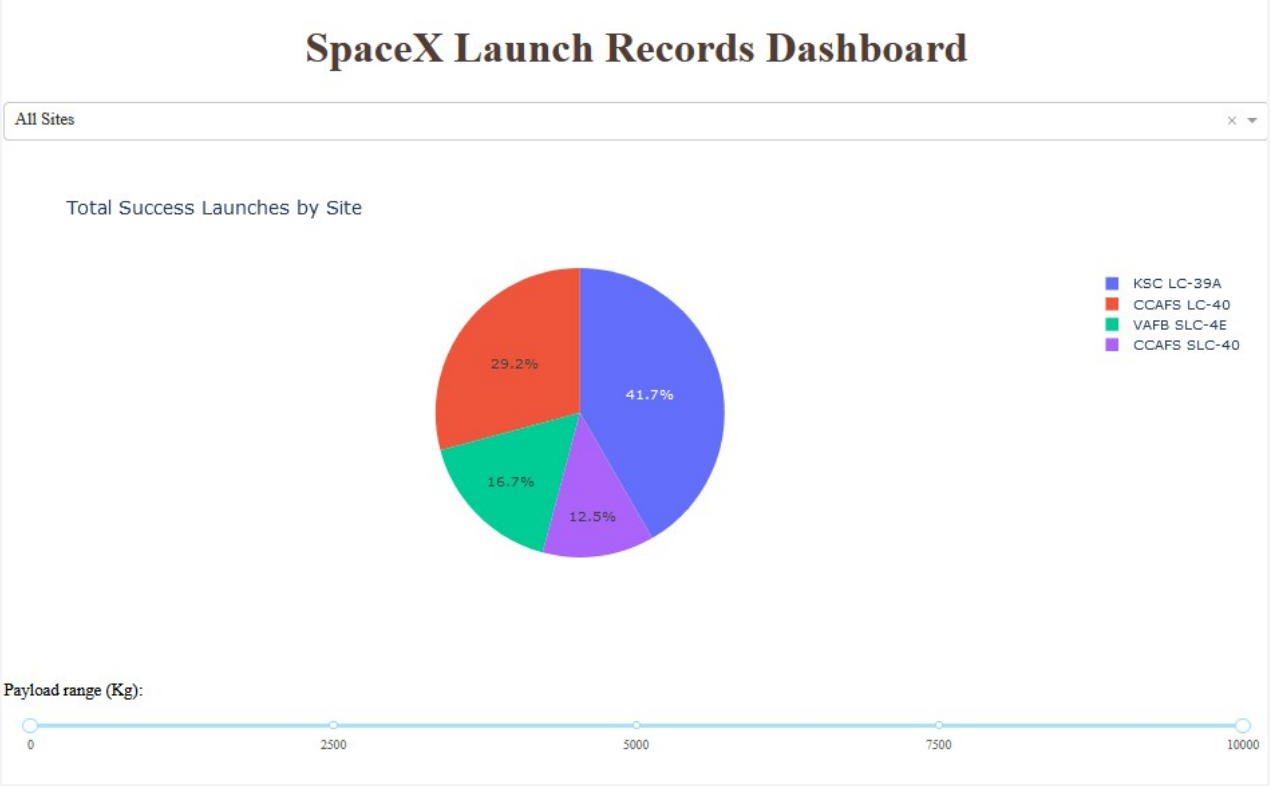
Sites must be far from cities, highways, or railways to minimize damage risks from failed launches.


Section 4
Build a Dashboard with Plotly Dash

Launch Successful Rate

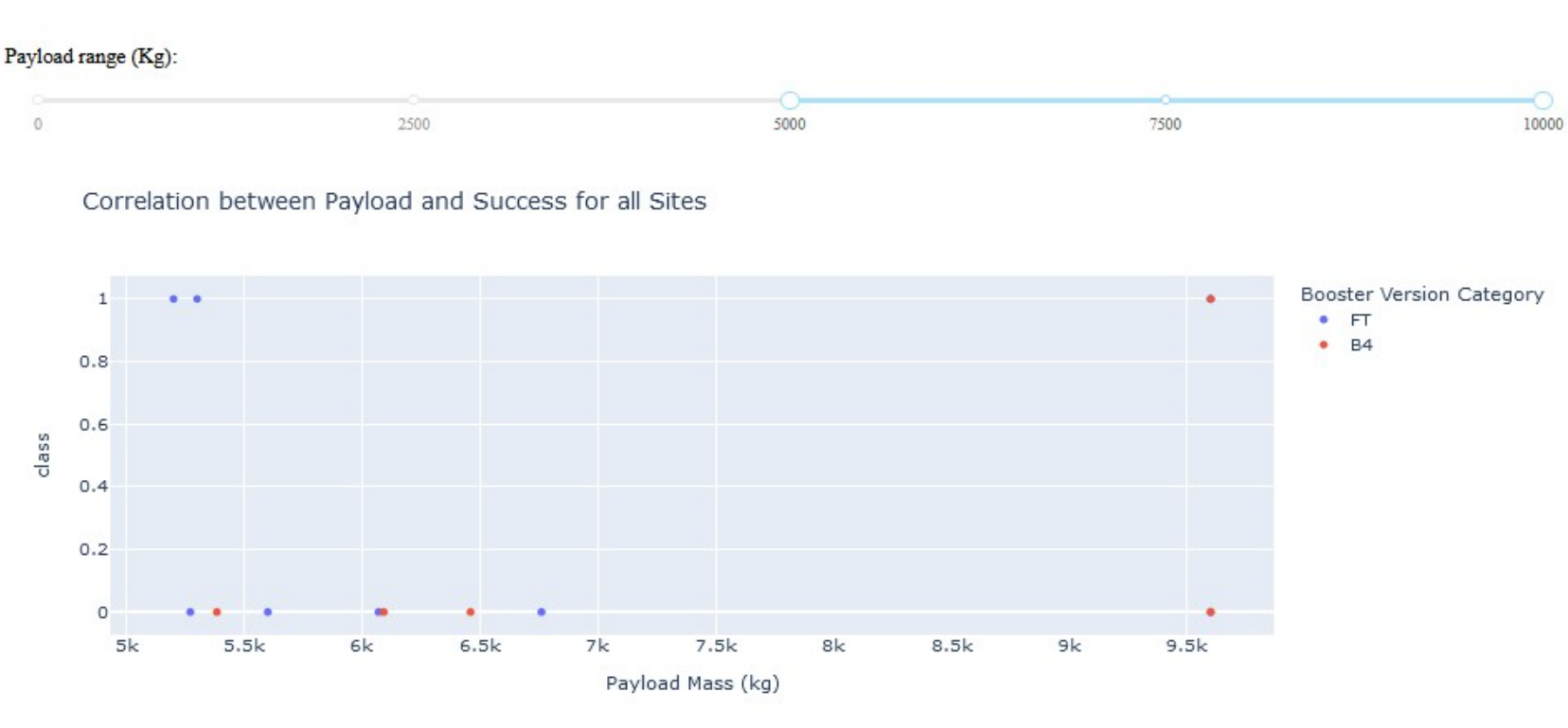
It is proved that KSC LC-39A is the most successful launch site.

CCAFS SLC-40 is the least successful.



Heavy Payloads Unsuccessful

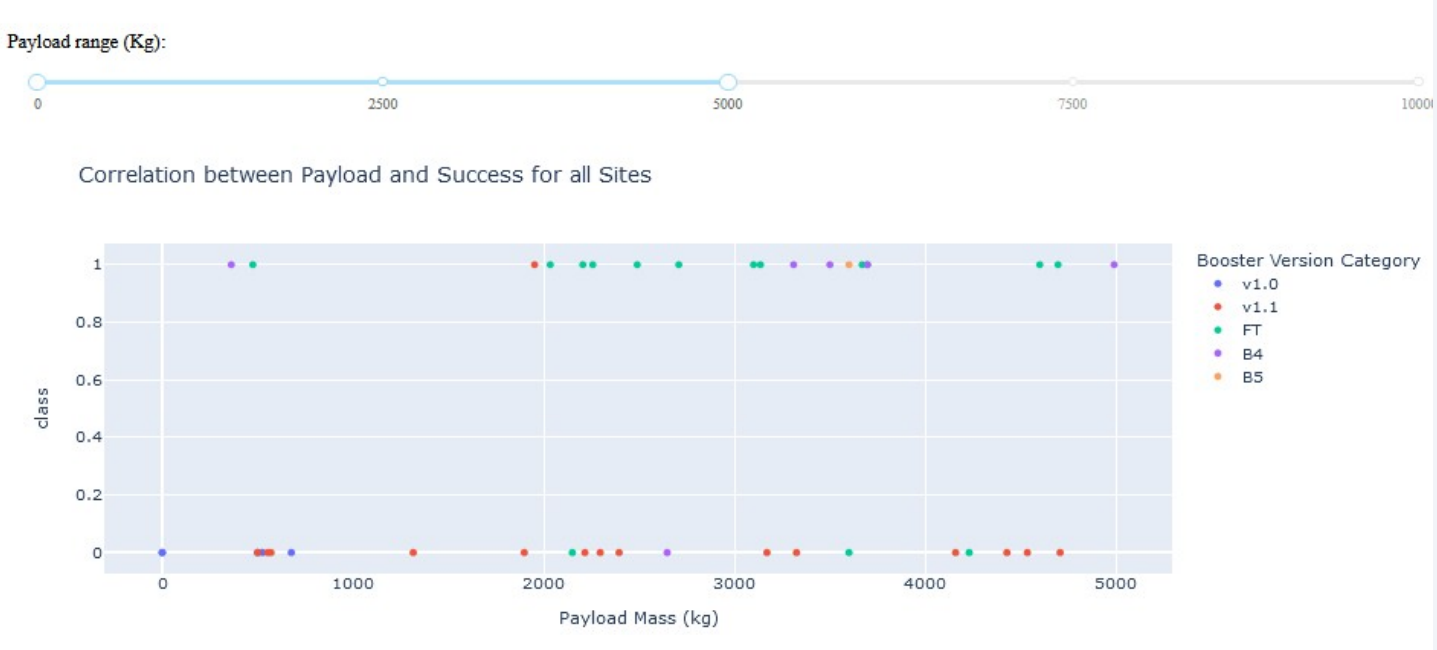
Payloads with 5000+ Kg mostly unsuccessful



Mid and Light Weight Payloads are Successful

From 5000 downwards, success rate drastically increases, as long as weigh is also higher than 2000.

If payload weight is less that 2000, we are in a similar situation to high weight payloads.



Section 5

Predictive Analysis (Classification)

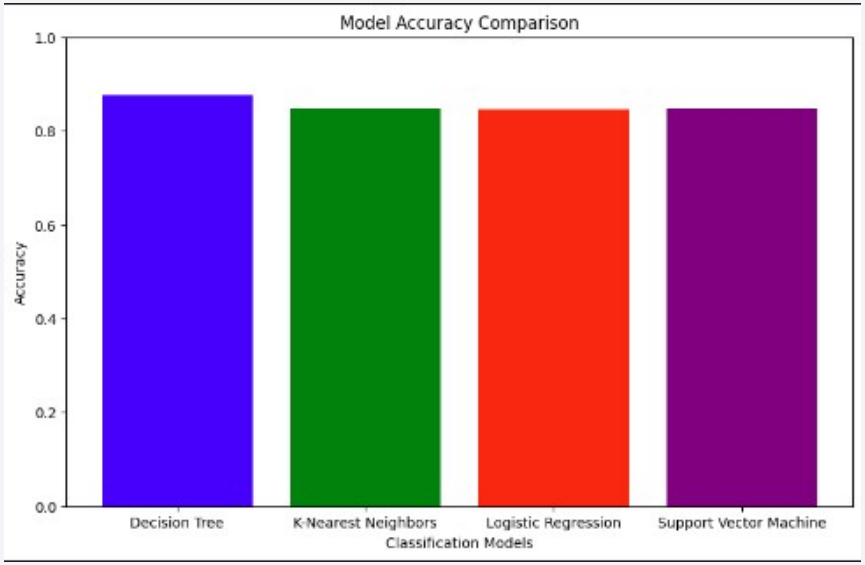
Classification Accuracy

Find the method performs best:

```
model_dic = {'KNeighbors':knn_cv.best_score_,
             'DecisionTree':tree_cv.best_score_,
             'LogisticRegression':logreg_cv.best_score_,
             'SupportVector': svm_cv.best_score_}

bestalgorithm = max(model_dic, key=model_dic.get)

print('Best model is', bestalgorithm,'with a score of', model_dic[bestalgorithm])

cv_objects = {
    'DecisionTree': tree_cv,
    'KNeighbors': knn_cv,
    'LogisticRegression': logreg_cv,
    'SupportVector': svm_cv
}

print('Best params is :', cv_objects[bestalgorithm].best_params_)
```

✓ 0.0s

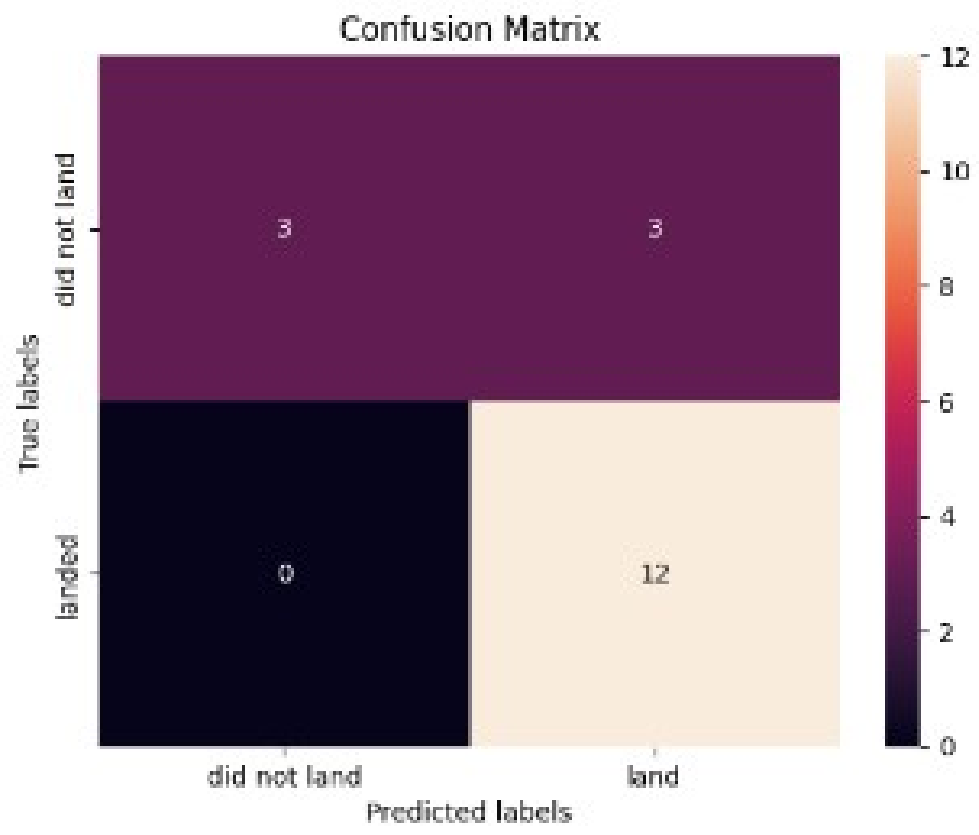
```
Best model is DecisionTree with a score of 0.875
Best params is : {'criterion': 'entropy', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}
```



By comparing all models, decision tree is slightly more accurate than the rest for this specific case.

All models are almost equally accurate.

Confusion Matrix



Confusion Matrix Output:

12 True positives

3 True negatives

3 False positives

0 False negatives

Conclusions

Model Performance: All models performed similarly on the test set, with the decision tree model achieving slightly better results.

Equatorial Advantage: Most launch sites are near the equator, leveraging Earth's rotational speed for a natural boost. This reduces fuel and booster requirements, lowering launch costs.

Coastal Locations: All launch sites are close to the coast, ensuring safety by minimizing risks to populated areas from debris or failed launches.

Launch Success Trends: Success rates have consistently improved over time.

KSC LC-39A Highlights:

- Highest overall success rate among launch sites.
- Achieved a 100% success rate for launches with payloads under 5,500 kg.

Orbit Success Rates: Orbits ES-L1, GEO, HEO, and SSO maintain a 100% success rate.

Payload Mass Influence: Across all launch sites, higher payload masses correlate with higher success rates.

Appendix

Thank you!