IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Richard Paredes

11/25/2024

Outline

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

Executive Summary

- Objective
  - Analyze SpaceX launch data using the REST API to explore trends, create visualizations, and develop a machine learning model to predict launch success.
- Key Insights
  - KSC LC 39A had the highest success rate (76.9%).
  - Launches with payloads between 2000 kg and 6000 kg are more likely to succeed.
  - Success rates have steadily increased since 2013.
  - The Decision Tree Classifier was the most accurate predictive model.
- Conclusion
  - The project identified critical factors influencing SpaceX launches and developed a reliable predictive model to optimize decision-making for future missions.

Introduction

- This project analyzes data from SpaceX's REST API to explore factors affecting launch success. The goal is to predict the likelihood of success based on variables like launch site, payload mass, and orbit type using machine learning models.
- SpaceX's numerous launches involve various factors that can influence mission success. By analyzing historical data, we aim to uncover patterns that help optimize future operations and decision-making.
- Launch site characteristics, payload mass, and orbital types all play a role in launch outcomes. This project examines these variables to understand how they affect success rates and identify trends over time.
- Questions to address: Which factors impact launch success? Which launch sites perform best? How do payload and orbit affect success? How has success evolved over time? Can machine learning predict success?
- This project aims to provide insights that optimize SpaceX's operations and improve decision-making for future launches.

Section 1

Methodology

Methodology

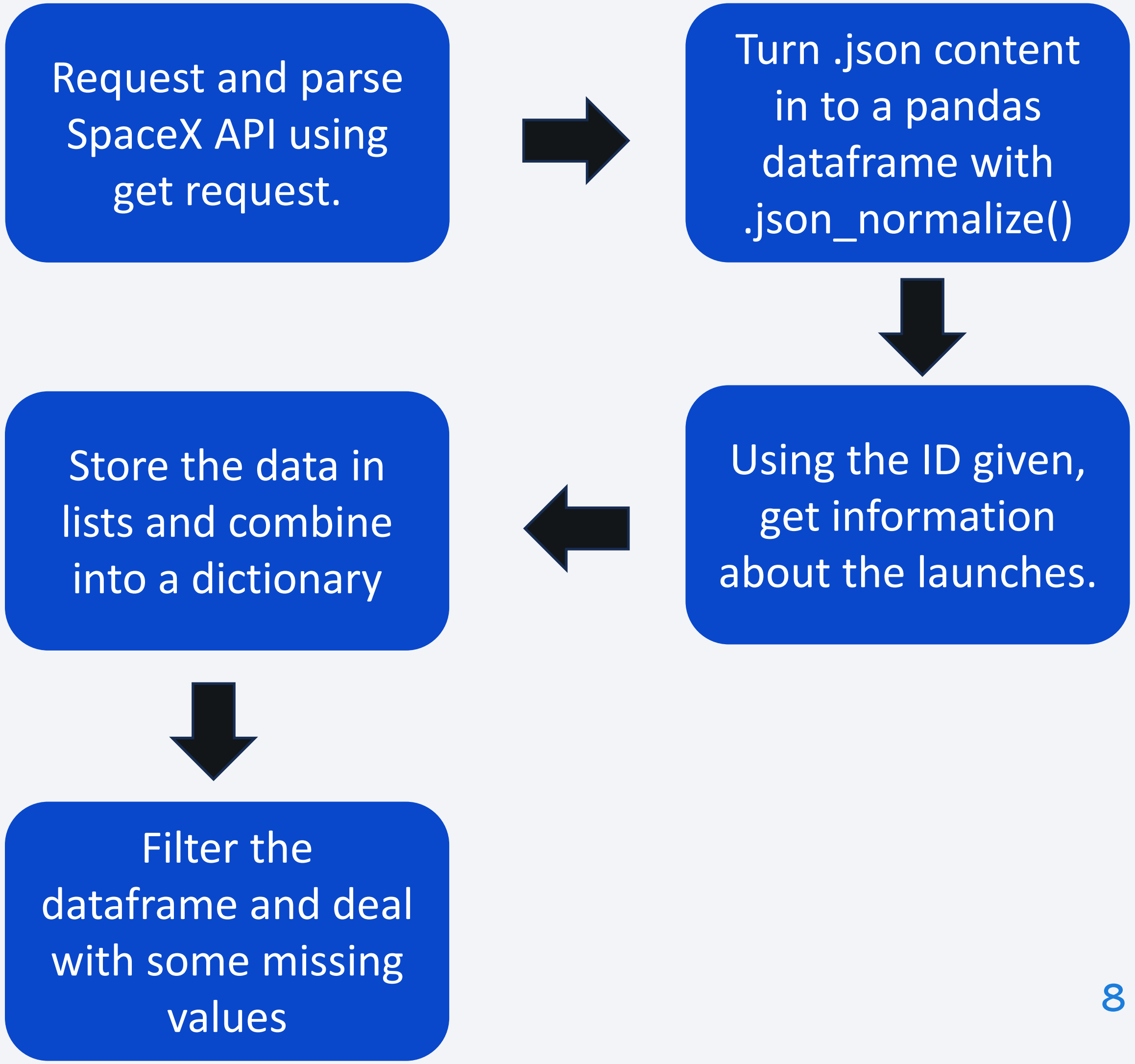
- Data collection methodology:
  - Data was collected from these sources:
    - Space X API - https://api.spacexdata.com/
    - List of Falcon 9 and Falcon Heavy launches - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - First, it was performed a data analysis. Then, a column named “Class” was created to label if a landing was successful or not.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was standardized and split into testing and training data. By finding the best hyperparameter, the method that performs best was found.

Data Collection

- Data sets were collected from the following sources:
  - Space X API - https://api.spacexdata.com/
  - List of Falcon 9 and Falcon Heavy launches - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

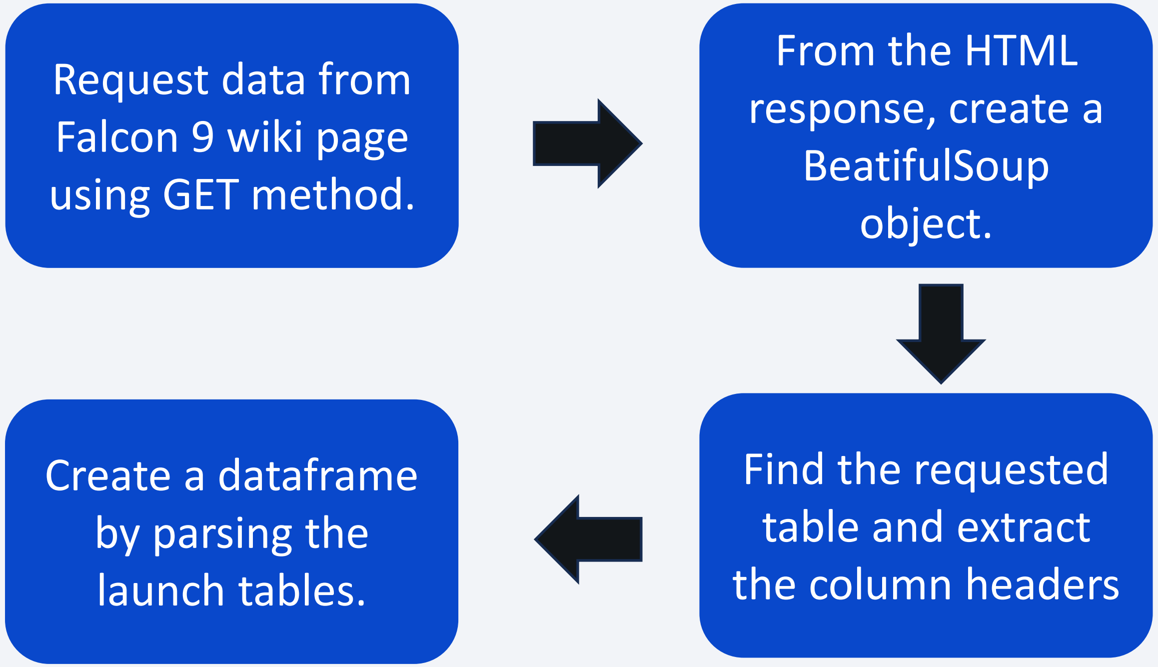
- Using the SpaceX API, a request was performed to get the dataset of the launches.
- After that, the requested data was cleaned.
- GitHub URL of the completed SpaceX API notebook: https://github.com/Richard-Paredes-1/DS-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

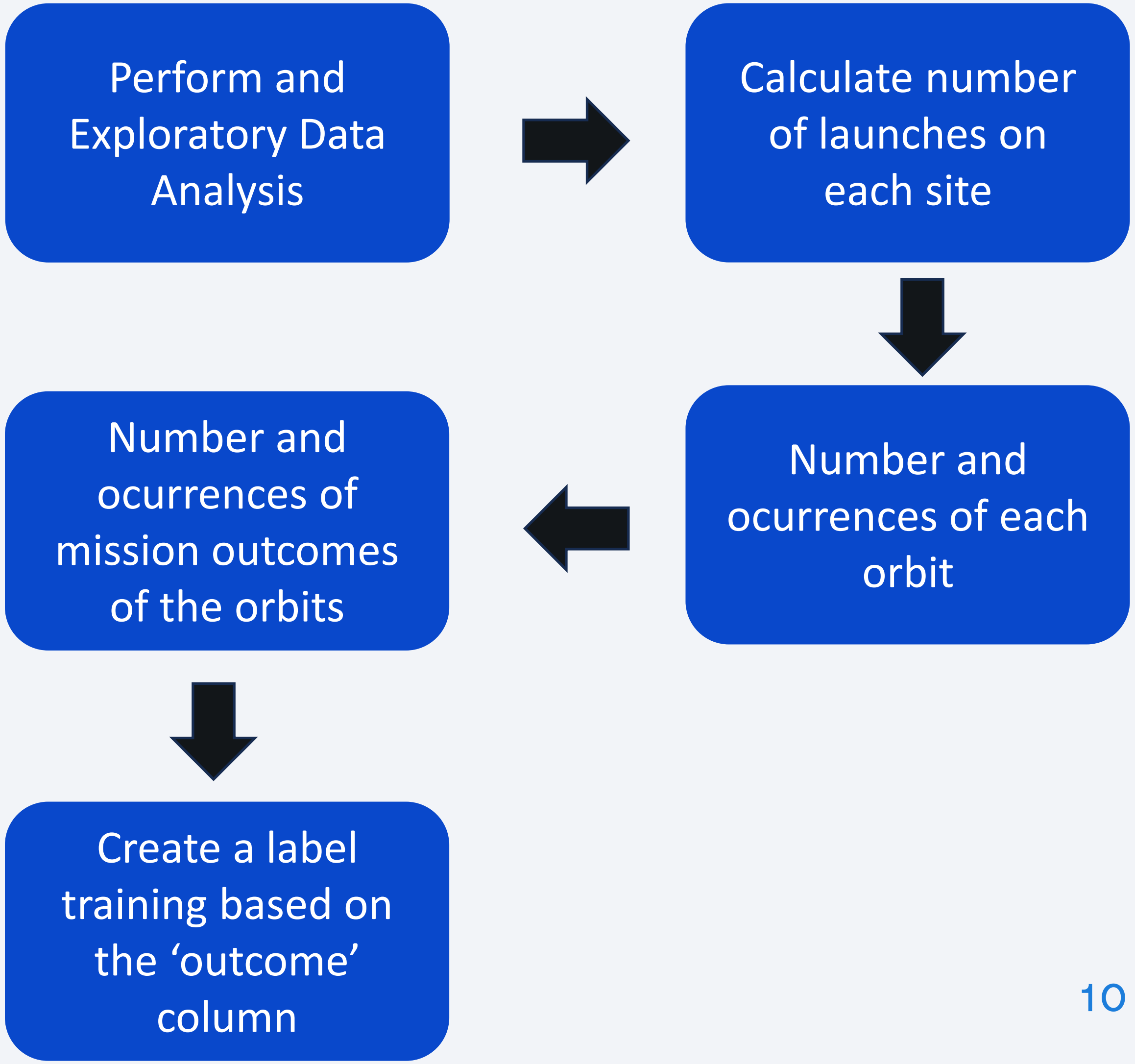
- Data from Falcon 9 Wiki page was requested to create a data frame using web scraping tools.

- GitHub URL of the completed Web Scarping notebook: https://github.com/Richard-Paredes-1/DS-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- It was performed an Exploratory Data Analysis to determine the labels for training supervised models.
- GitHub URL of the completed Data Wrangling notebook: https://github.com/Richard-Paredes-1/DS-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- An Exploratory Data Analysis was conducted to better understand the data.
- It started with a scatterplot to examine the relationship between Flight Number and Launch Site, Payload Mass and Launch Site, Flight Number and Orbit Type, and Payload Mass and Orbit Type.
- A bar chart was created to visualize the success rate of each orbit type.
- A line chart was designed to determine the success rate for each year.
- GitHub URL of the completed EDA with data visualization notebook: https://github.com/Richard-Paredes-1/DS-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Identify and display the distinct launch sites used in the space missions.
- Calculate and display the total payload mass transported by boosters launched under NASA's CRS missions.
- Determine and display the date of the first successful landing on a ground pad.
- Retrieve and display the names of boosters that achieved successful landings on a drone ship, with payload masses between 4000 and 6000.
- Count and display the total number of missions with successful and failed outcomes.
- Identify and display the booster versions that carried the highest payload mass.
- Extract and display records showing month names, failed drone ship landing outcomes, booster versions, and launch sites for missions conducted in 2015.
- Rank the frequency of different landing outcomes (e.g., drone ship failures or ground pad successes) between June 4, 2010, and March 20, 2017.
- GitHub URL of the completed EDA with SQL notebook: https://github.com/Richard-Paredes-1/DS-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

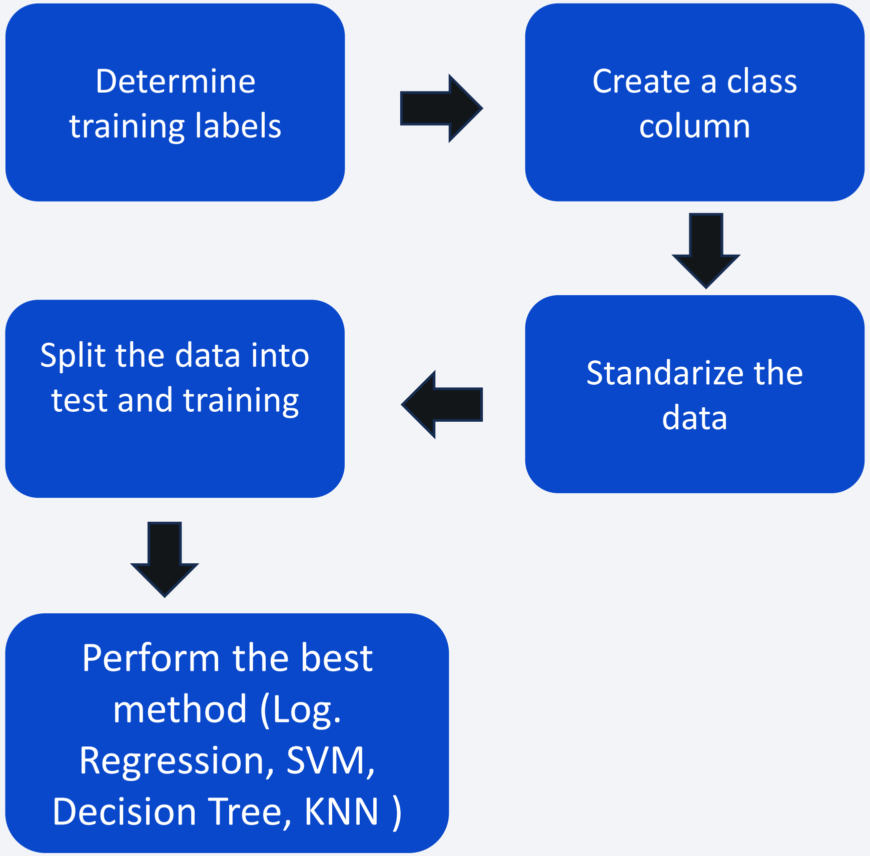
- It was marked all launch sites on a map using Folium. Created and added folium.Circle and folium.Marker for each launch site on the site map.
- It was marked the success/failed launches for each site on the map. For each launch result in spacex_df data frame, it was added a folium.Marker to marker_cluster.
- It was calculated the distances between a launch site to its proximities. It was marked down a point on the closest coastline, highway, railway and city using MousePosition and Polyline and calculated the distance between the coastline, highway, railway and city point and the launch site.
- GitHub URL of the completed interactive map with Folium map: https://github.com/Richard-Paredes-1/DS-Capstone/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- It was added a dropdown list to enable Launch Site selection. The purpose of this was to interact with the visualizations that were added.
- It was added a pie chart to show the total successful launches count for all sites to see if there is any pattern between the successful launches and the launch sites.
- It was added to select payload range to interact with the scatterplot.
- It was added scatterplot to show the correlation between payload and launch success of each site.
- GitHub URL of the completed Plotly Dash lab: https://github.com/Richard-Paredes-1/DS-Capstone/blob/main/plotly%20visualization%20code.txt

Predictive Analysis (Classification)

- The train_test_split function was used to divide the data, X and Y, into training and testing sets. The test_size parameter was set to 0.2, and random_state was set to 2. The resulting training and testing datasets were assigned to the following labels: X_train, X_test, Y_train, Y_test.
- Then, a logistic regression, support vector machine, decision tree classifier and k nearest neighbor were performed to compare the accuracy of each method and plot a confusion matrix.
- After that, it was obtained the best method for the model.
- GitHub URL of the completed predictive analysis lab: https://github.com/Richard-Paredes-1/DS-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

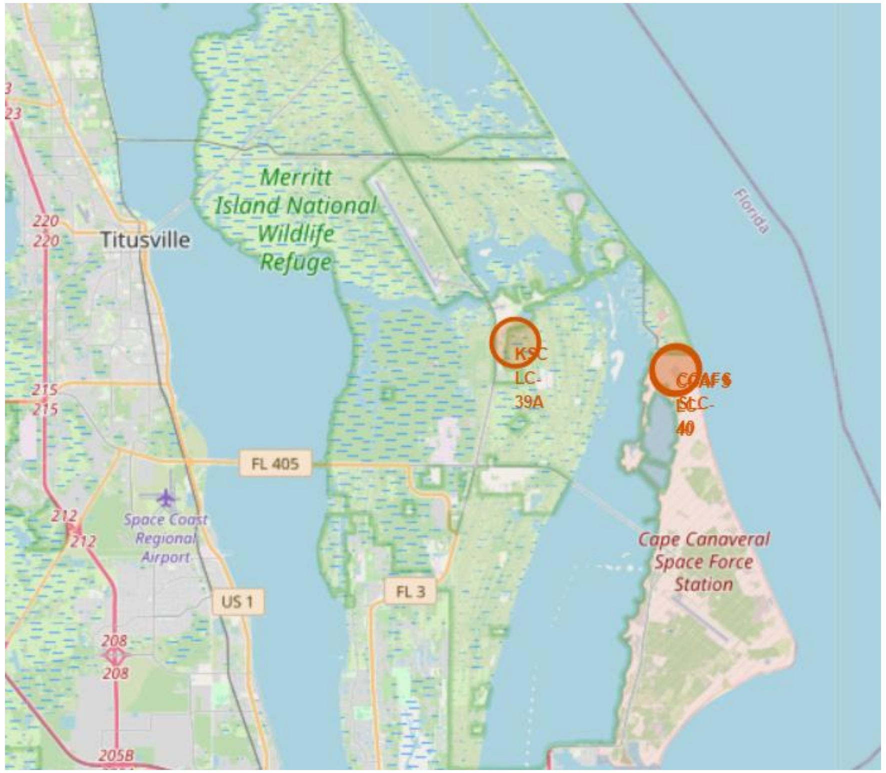


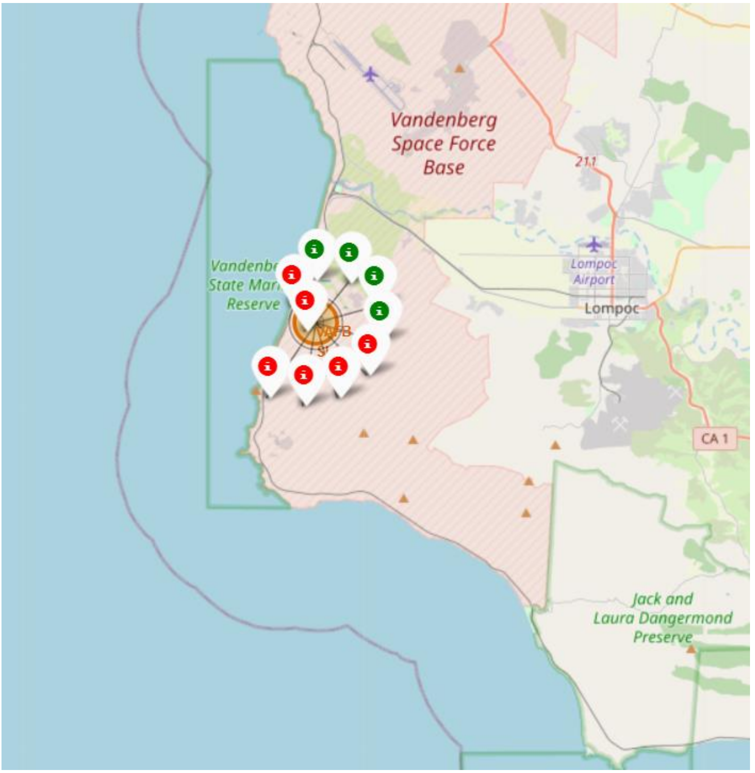
Results

- Exploratory data analysis results
  - Space X have 4 launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40.
  - Most of the launch sites are in CCAFS SLC40.
  - Most of the launches are less than 6000 kg.
  - The orbits with a 100% of success landings were ES-L1, GEO, HEO and SSO.
  - Success rate of landings consistently improved from 2013 to 2020.
  - There were 12 boosters which have carried 15,600 kg (maximum payload mass)
  - Date of the first successful landing outcome on ground pad: 2015-12-22.
  - Total payload carried by boosters from NASA is 2,534.6 kg.
  - The two failed landing outcomes occurred in January and April of 2015.

Results

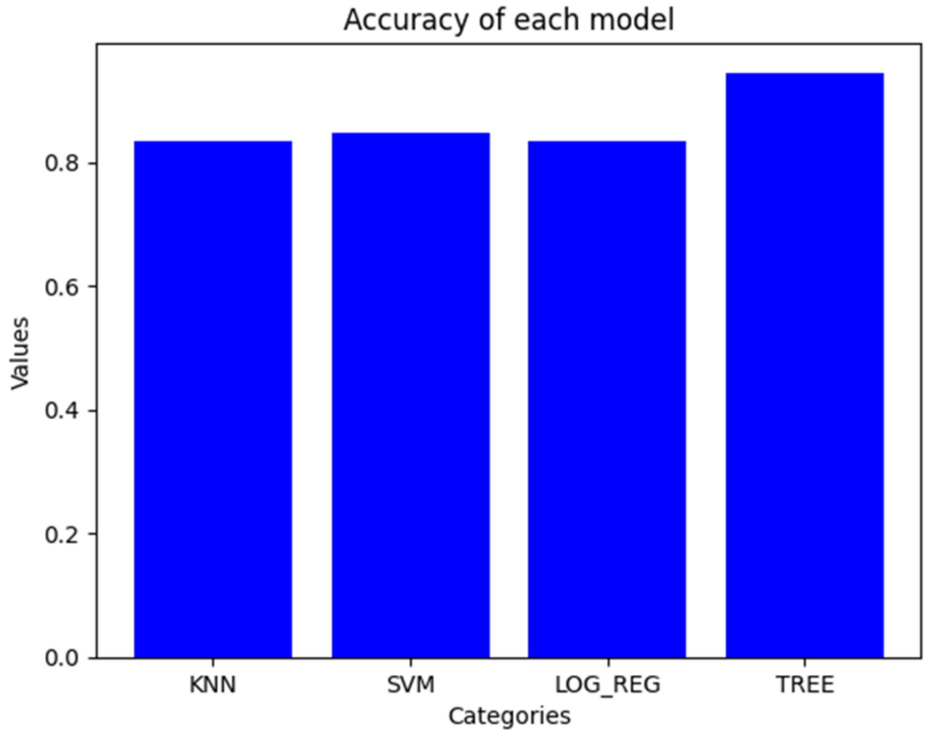
- Interactive analytics demo shows that the launch sites appear to be strategically positioned, as they are located near railways, highways, cities, and coastlines. This placement may be connected to considerations such as safety analysis, impact range, or the tracking of space launches.





Results

- The bar chart shows the accuracy of four different models: KNN, SVM, Logistic Regression, and Decision Tree. Among them, the Decision Tree classifier achieved the highest accuracy, indicating it outperformed the other models in this evaluation.

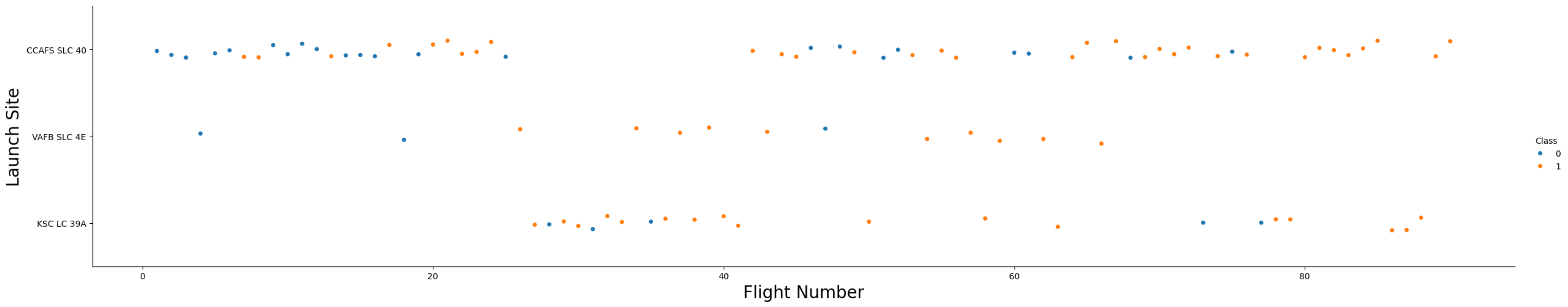


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

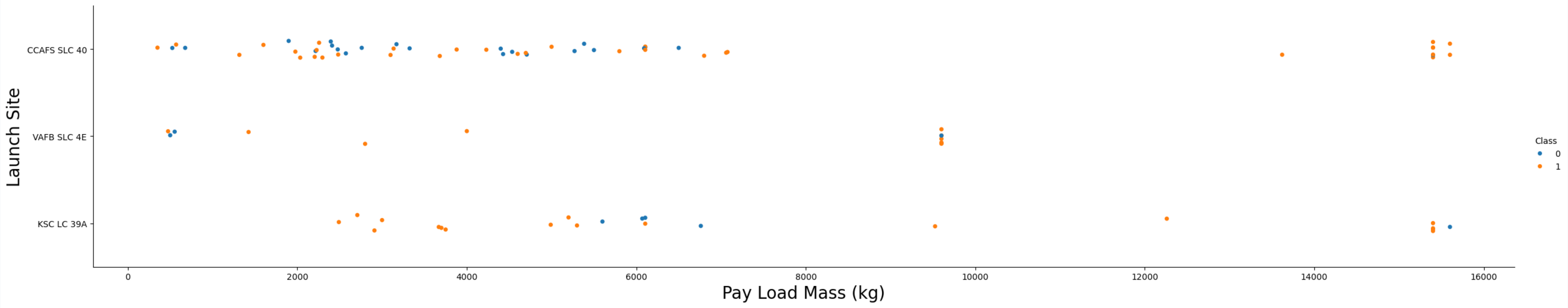
- Scatter plot of Flight Number vs. Launch Site



It can be observed that most of the launch sites are in CCAFS SLC 40, especially during the early flight numbers.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

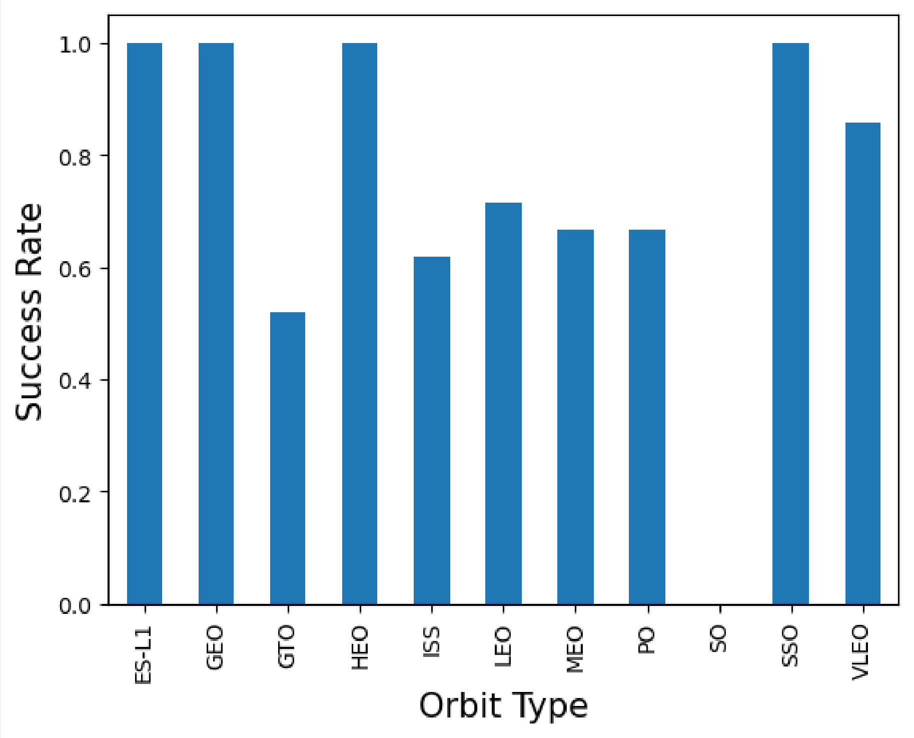


It can be observed that most of the payload masses, which fall within a range of up to 8000 kg, are launched from CCAFS SLC 40.

Success Rate vs. Orbit Type

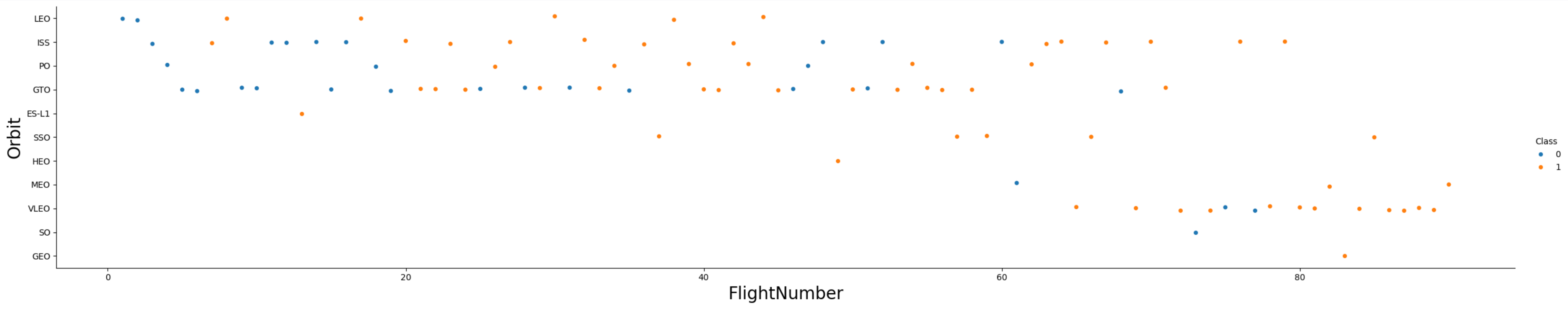
- Bar chart for the success rate of each orbit type.

It can be observed that the orbits with a 100% of success were ES-L1, GEO, HEO and SSO.



Flight Number vs. Orbit Type

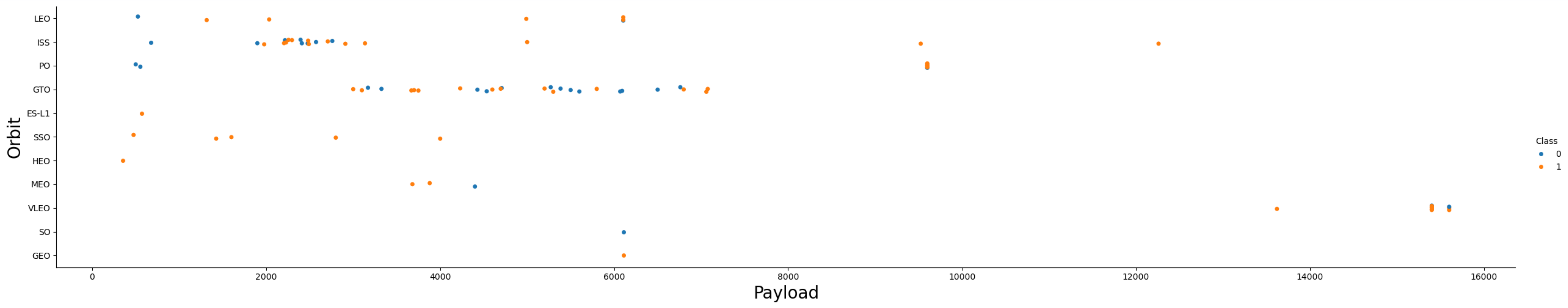
- Scatter plot of Flight number vs. Orbit type



The main pattern is that the majority of of flights occurred on orbit GTO, PO, ISS, and LEO. Nevertheless, this changed after the flight number 60.

Payload vs. Orbit Type

- Scatter plot of payload vs. orbit type

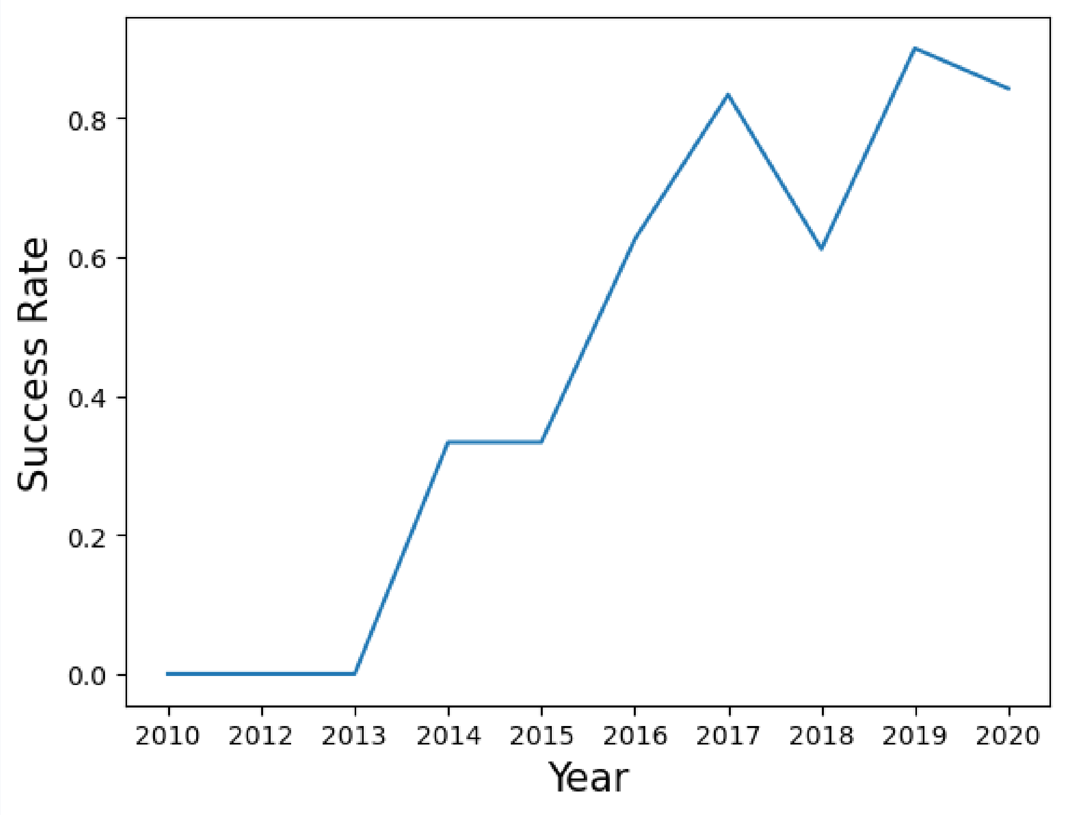


It is observed that most flights were once again carried out with a payload mass of less than 8000 kg, and the majority were conducted in orbits GTO and ISS.

Launch Success Yearly Trend

- Line chart of yearly average success rate

It can be observed that the success rate consistently improved from 2013 to 2020.



All Launch Site Names

- Names of the unique launch sites.
- The launch sites are
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Task 1

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

```
 * 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'

- 5 records where launch sites begin with `CCA`
- All of them are CCAFS LC-40, F9 v1.0 and the mission outcome was a success.

Task 2

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_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

- Total payload carried by boosters from NASA.

Task 3

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

```
[13]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

[13]:

SUM(PAYLOAD_MASS__KG_)
45596

- The total is 45,596 kg.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

Task 4

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- The average payload mass is 2,534.6 kg.

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(date)
2015-12-22

- It happened on 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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, landing_outcome, payload_mass__kg_ FROM SPACEXTABLE WHERE landing_outcome = 'Success (drone ship)' AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
```

```
 * 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

- The list name is F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, COUNT(mission_outcome) AS TOTAL_NUMBER FROM SPACEXTABLE GROUP BY mission_outcome;
```

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

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

- Successful missions are 100.
- There is 1 failed mission.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery ¶

```
%sql SELECT booster_version, payload_mass__kg_ FROM SPACEXTABLE WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXTABLE)
```

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

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

- There are 12 boosters which have carried 15,600 kg (maximum payload mass)

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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.

```
1 SELECT substr(Date, 6,2) AS Months, landing_outcome, booster_version, launch_site FROM SPACEXTABLE WHERE landing_outcome = 'Failure (drone ship)' AND substr(Date,0,5) = '2015'
```

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

Months	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

- The two failed landing outcomes occurred in January and April of 2015.

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

- 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

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 RANKING FROM SPACEXTABLE WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY RANKING DESC
```

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

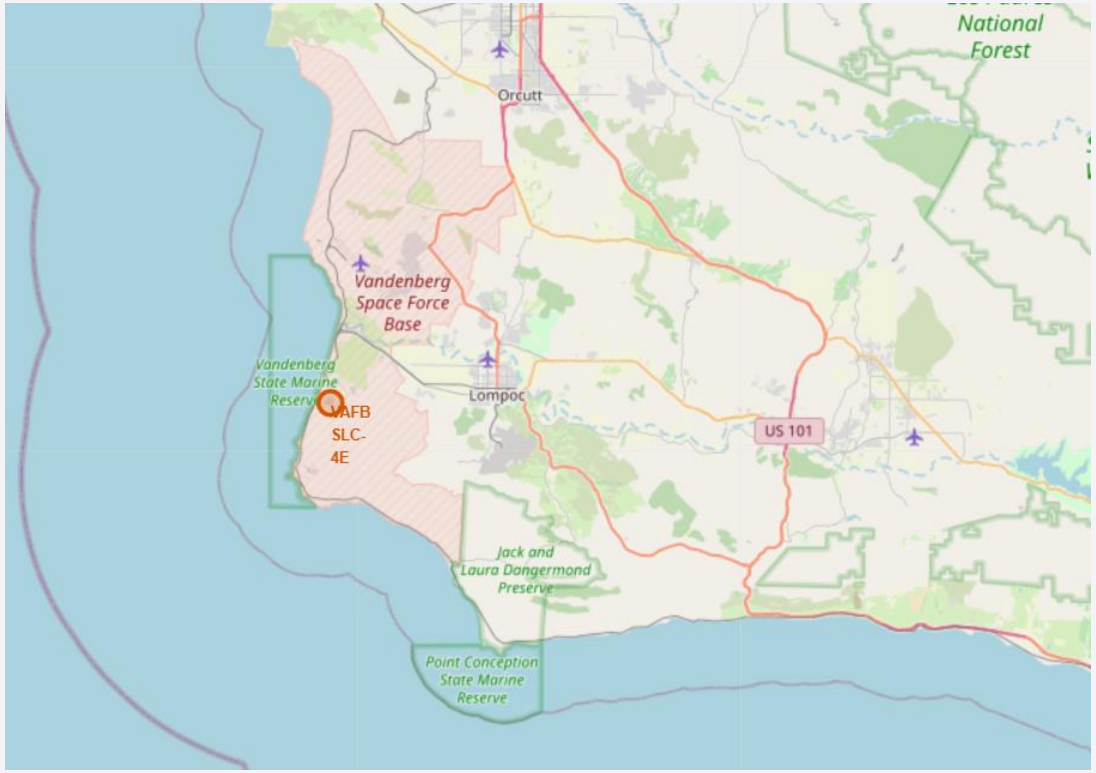
Landing_Outcome	RANKING
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

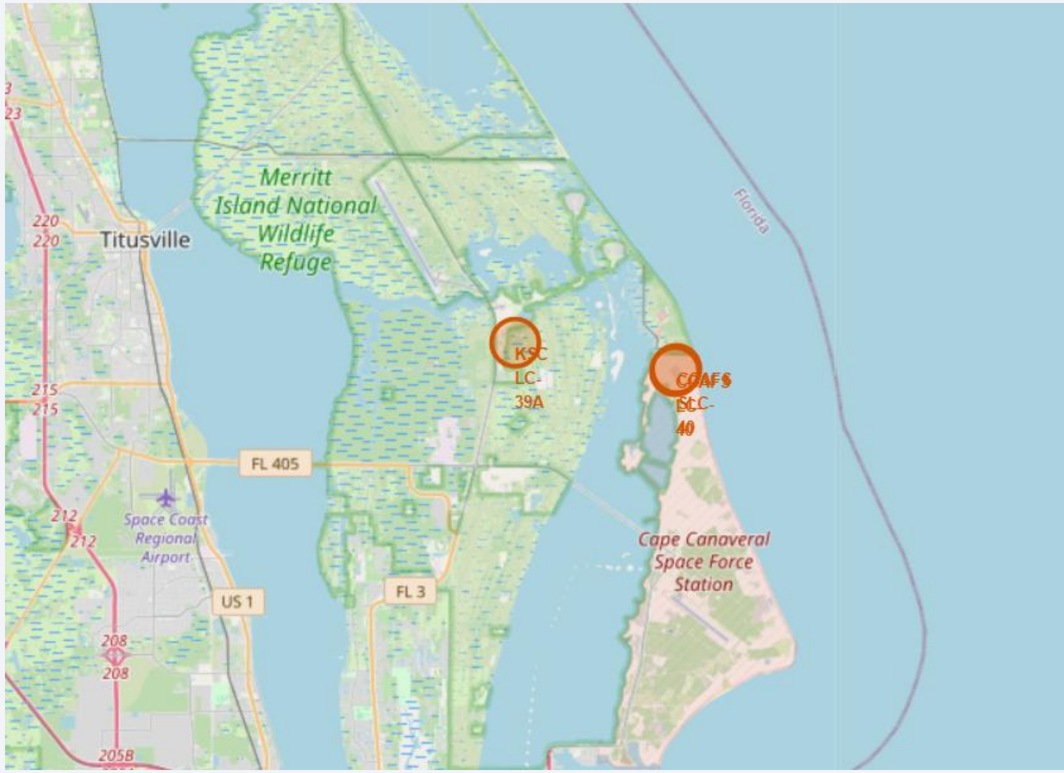
The main finding is that there were 10 landing outcome with no attempts. And 5 landing outcome with drone ship successes and 5 failures.

Section 3

Launch Sites Proximities Analysis

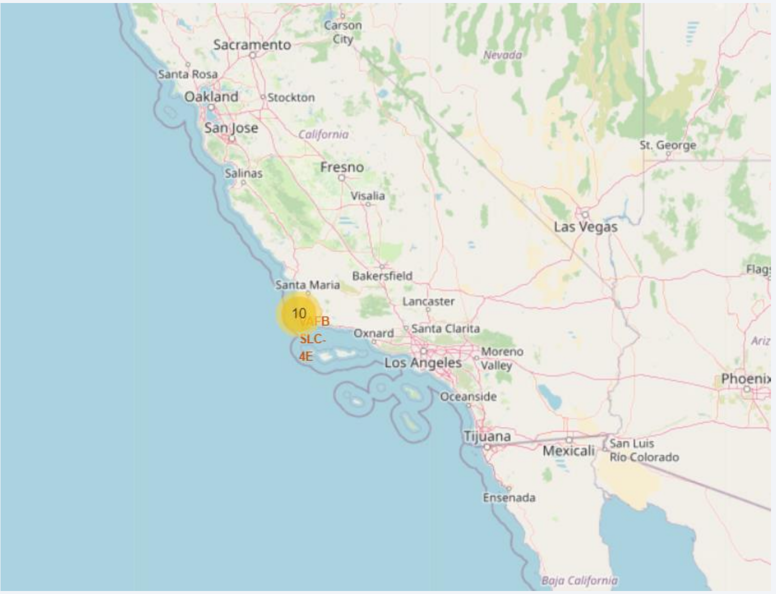
All launch sites on a map

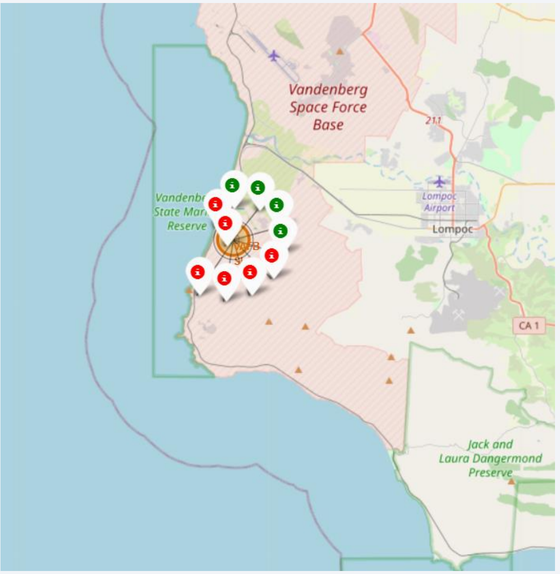




- The launch sites are in California and Florida.

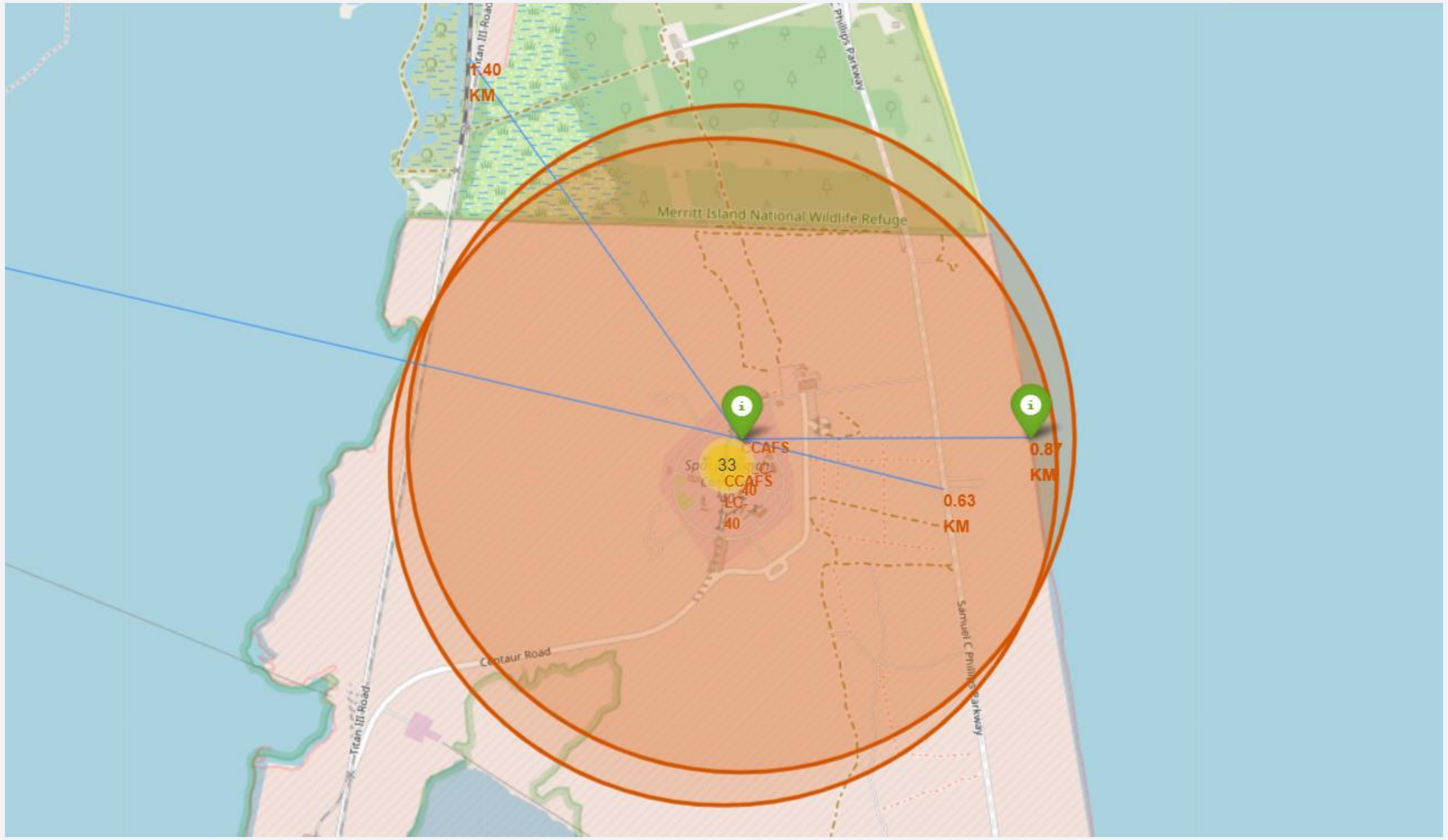
Launch sites markers and their outcomes





- The number of missions carried out at each launch site was obtained. Additionally, the record of whether the mission was successful or failed is observed. It is noted that most of the tests were conducted in Florida.

Coverage and Proximity Areas of Launch Sites

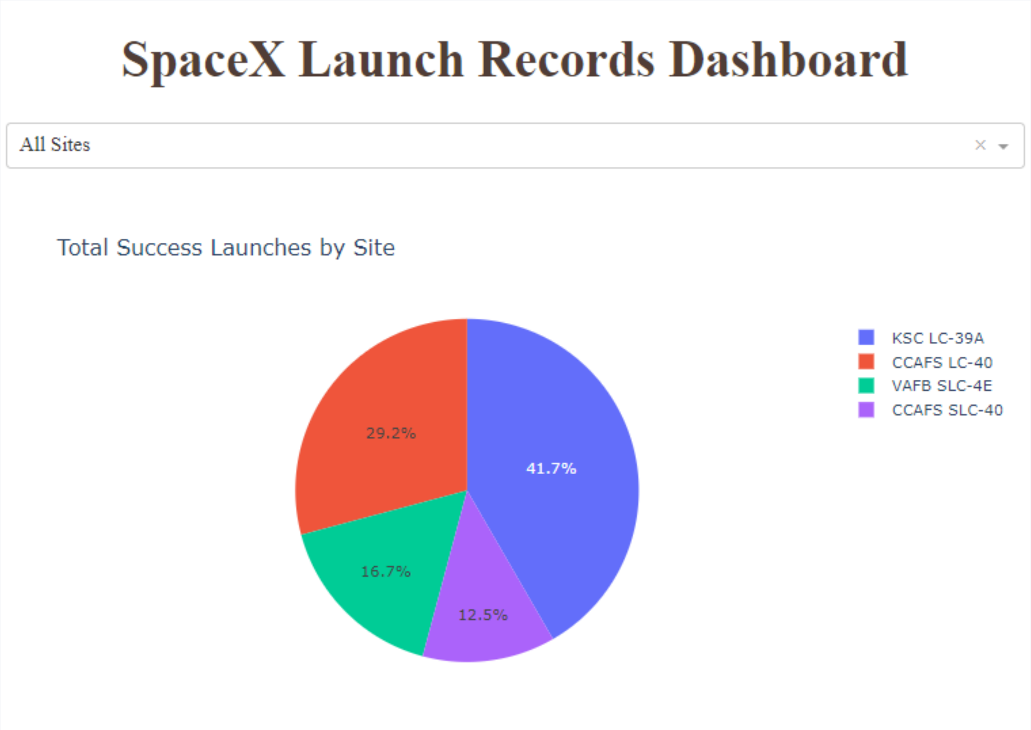


- It seems that the launch sites are characterized by being in strategic locations, as they are quite close to the railway, highway, city, and coastline. This could be related to safety analysis, impact range, or tracking of space launches.

Section 4

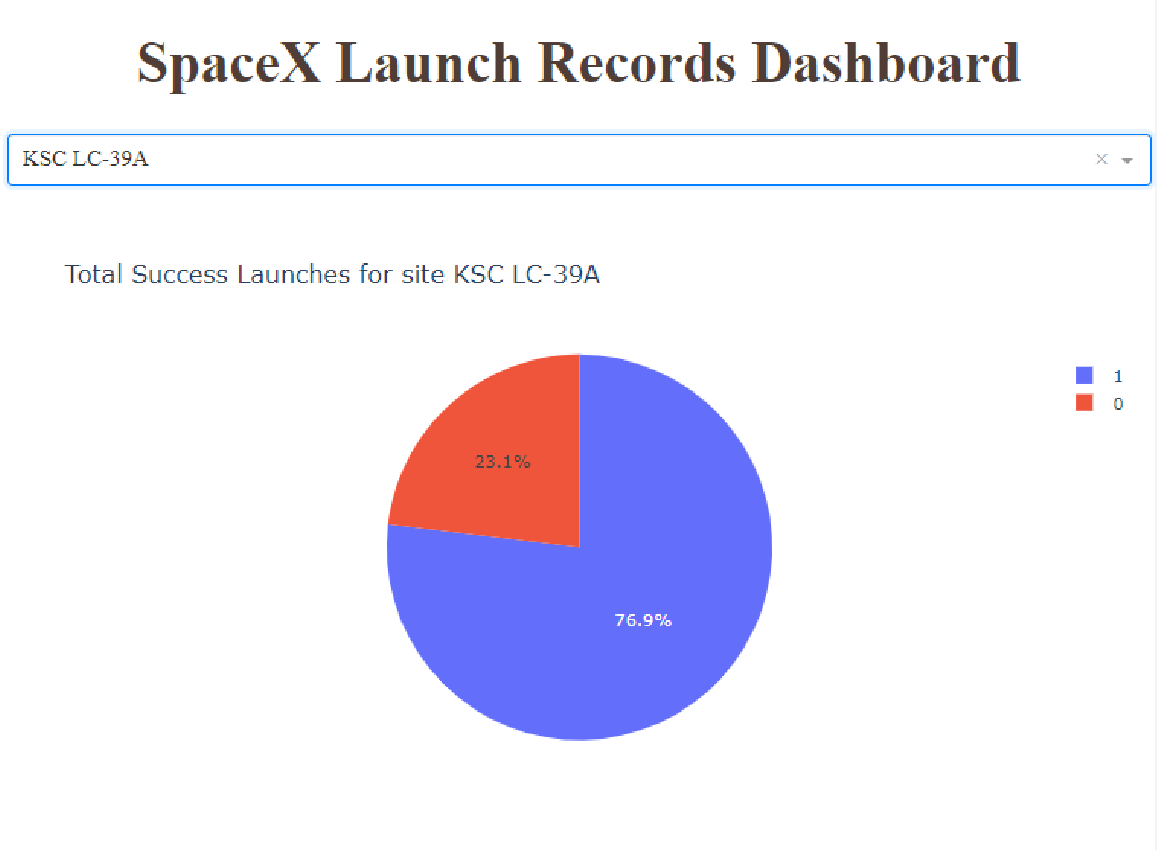
Build a Dashboard with Plotly Dash

Total Success Launches by Site



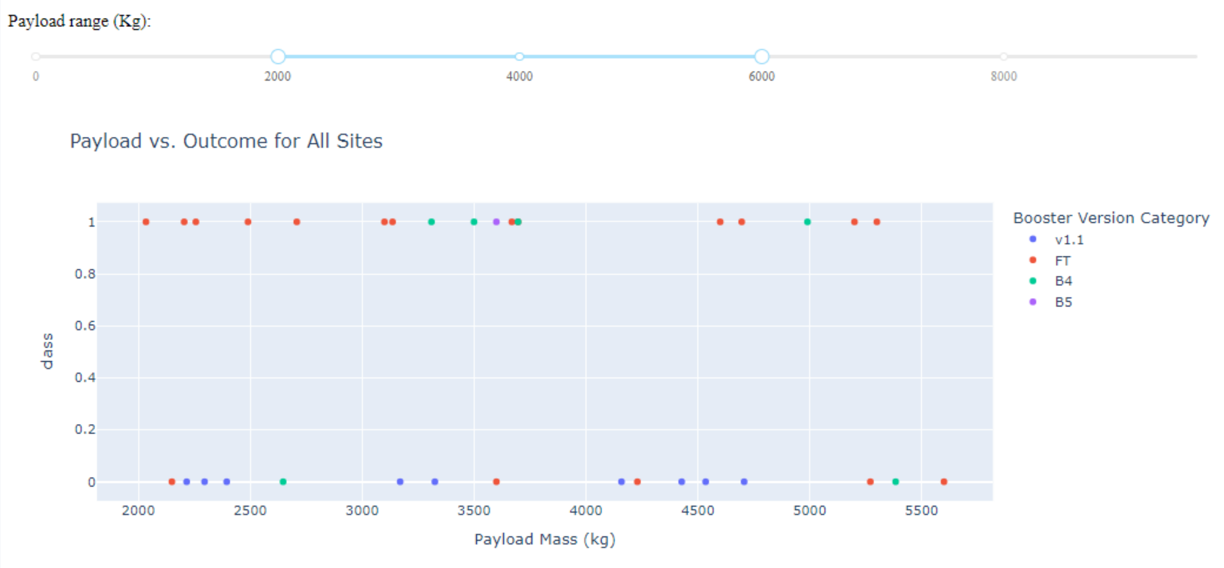
- It seems that successful launches and the launch sites might be correlated.

The most successful launch site



- It is possible to say that the launch with the higher success ratio is KSC LC-39A.

<Dashboard Screenshot 3>



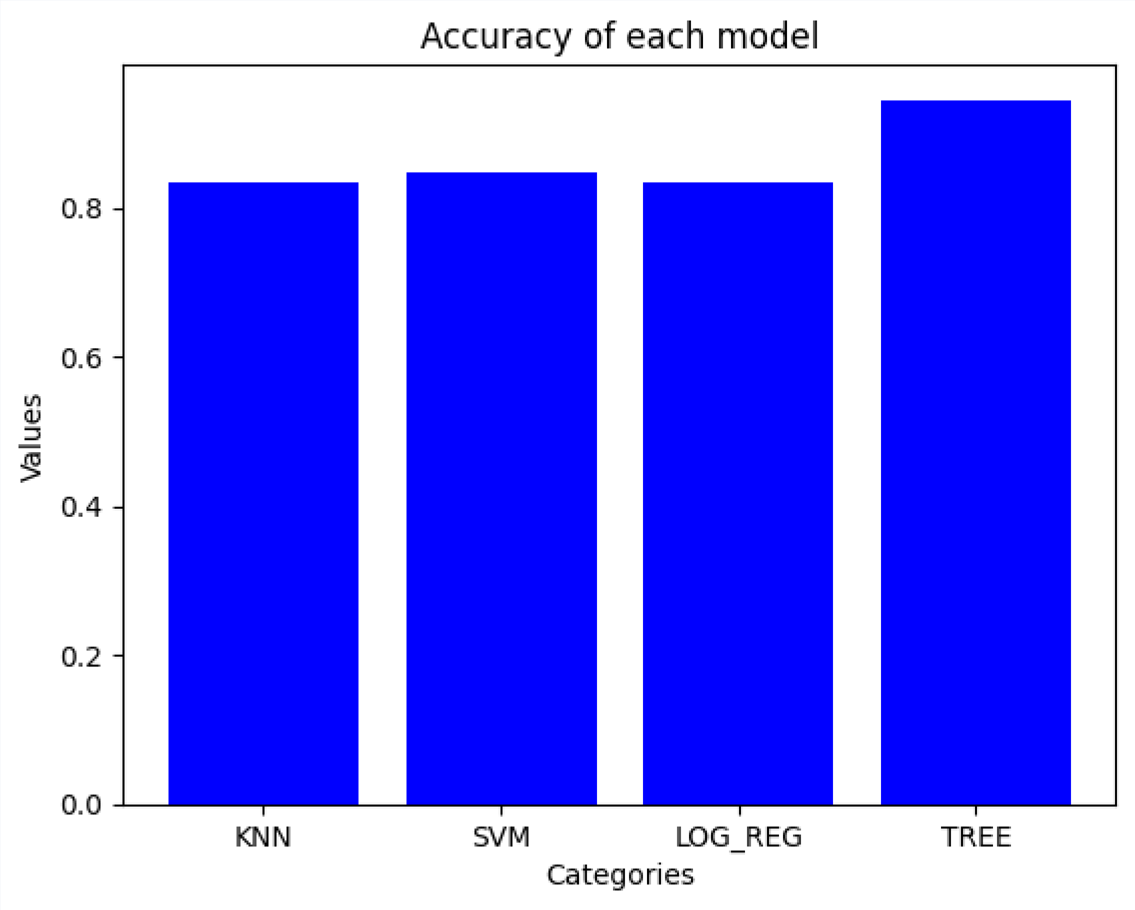
- The majority of launches with a payload mass between 2000 kg and 6000 kg tend to have more successful outcomes.

Section 5

Predictive Analysis (Classification)

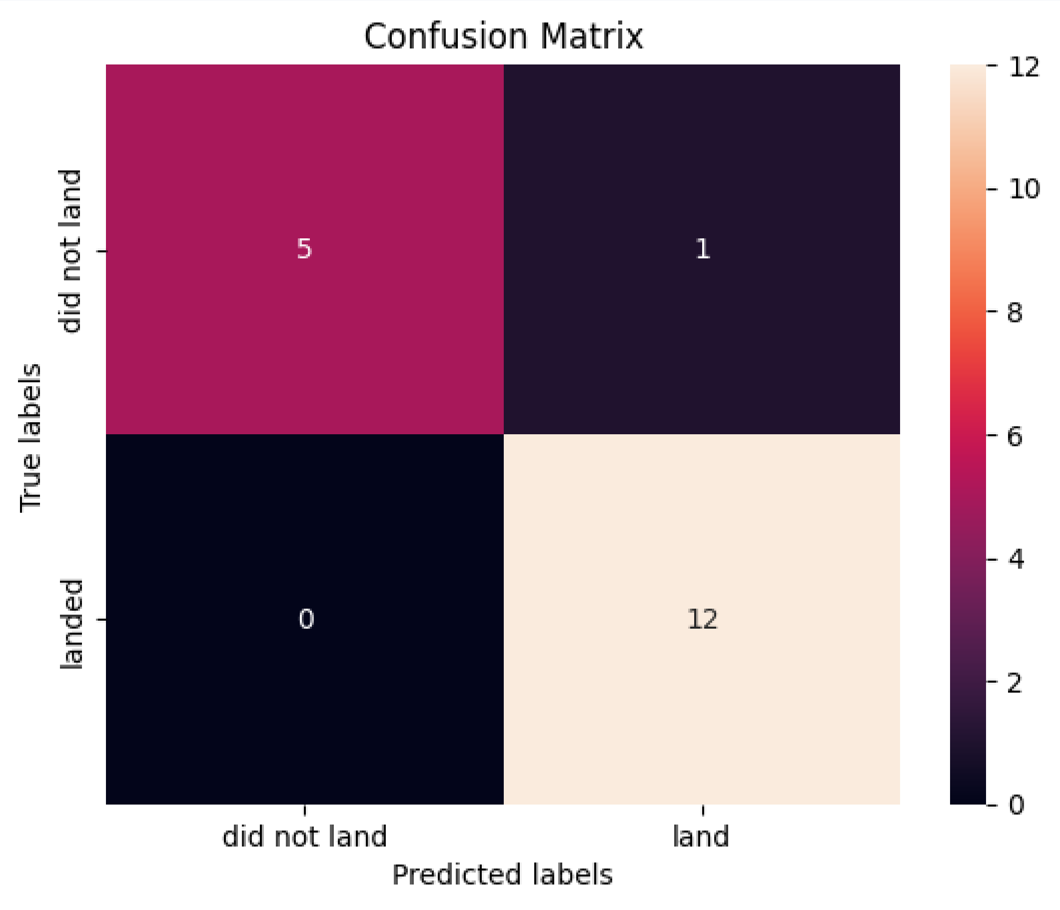
Classification Accuracy

- From the four models tested, the one with the highest accuracy was the decision tree classifier model.



Confusion Matrix

- The confusion matrix of the Decision Tree Classifier demonstrates its accuracy by presenting a significantly higher count of true positives and true negatives compared to false positives and false negatives.



Conclusions

- An Exploratory Data Analysis was performed to understand the data.
- The analysis revealed that there are four main launch sites, with KSC LC 39A standing out for its top success rate of 76.9%.
- The data also indicated a steady increase in flight success rates beginning in 2013.
- Furthermore, the placement of the launch sites appears to be strategically chosen to ensure proper safety analysis, impact range, and launch tracking.
- Most launches with a payload mass ranging from 2000 kg to 6000 kg were found to have higher success rates.
- Of the four models evaluated, the Decision Tree Classifier demonstrated the best performance with the highest accuracy.

Appendix

Thank you!