# Data Science Capstone Project

Lakshay Garg

Aug 31, 2024

# Outline:

- Executive Overview
- Project Introduction
- Methodological Approach
- Results & Analysis
- Conclusions & Insights
- Supplementary Information

# Executive Summary

**Summary of Approaches:**

- **Data Gathering:** Utilized SpaceX REST API and web scraping from Wikipedia.
- **Data Preparation:** Conducted data cleaning and encoding for binary classification.
- **Exploratory Analysis:** Employed data visualization and SQL queries to uncover insights.
- **Interactive Analytics:** Created dynamic maps with Folium and dashboards with Plotly Dash.
- **Predictive Modeling:** Developed and evaluated classification models.

**Summary of Outcomes:**

- Key findings from exploratory data analysis.
- Screenshots of interactive analytics tools.
- Results from predictive modeling.

# Introduction

**Background:** SpaceX has revolutionized space travel by drastically reducing launch costs through reusable rockets. This project aims to predict the success of Falcon 9 first-stage landings using public data and machine learning models.

**Research Questions:**

- How do factors like payload mass and launch site affect landing success?
- Has the landing success rate improved over time?
- What is the most effective classification algorithm for this scenario?

# METHODOLOGIES

# Methodologies

**Data Collection:**

- **Sources:** SpaceX REST API, Wikipedia.
- **Process:** Acquired launch data, filtered for relevant information, and handled missing values.
- **Tools:** Employed API requests and web scraping techniques.

**Data Analysis:**

- **Exploratory Analysis:** Visual and SQL-based analysis.
- **Interactive Tools:** Built visualizations with Folium and dashboards with Plotly Dash.
- **Predictive Modeling:** Applied various classification models and evaluated their performance.

# Data Collection

API

**Request data**
SpaceX API

**Decode response**
.json_normalize()

**Create dictionary**
SpaceX data

**Create dataframe**
SpaceX dictionary

**Filter dataframe**
Falcon 9 launches

**Missing values**
mean()

**Export data**
csv

# Data Collection

## Web Scraping

**Request data**
wikipedia

**BeautifulSoup Obj**
html response

**Extract Col names**
Table header

**Collect data**
Parsing table

**Create dictionary**
Table data

**Create dataframe**
dictionary

**Export data**
csv

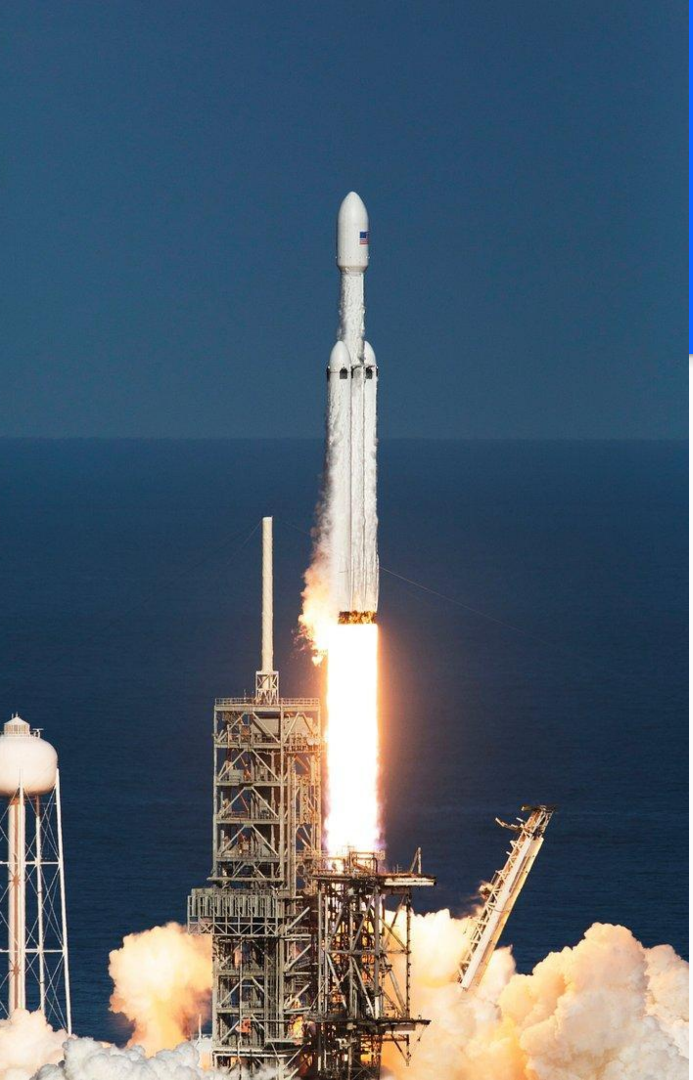
# Data Wrangling

**Data Cleaning:**

- Addressed missing values and standardized the data.
- Encoded categorical variables for classification.
- Exported the cleaned data for further analysis.

**Exploratory Data Analysis (EDA):**

- Analyzed launch site frequency, orbit types, and mission outcomes.
- Created training labels for machine learning models.



# Exploratory Data Analysis (EDA) with Visualization

**Key Visual Insights:**

- **Launch Site Analysis:** Examined success rates across different launch sites.
- **Payload Mass:** Analyzed the relationship between payload mass and landing success.
- **Orbit Type:** Assessed the success rate across different orbit types.
- **Yearly Trends:** Tracked the increase in success rates over time.

**Visualization Tools:**

- Scatter plots, bar charts, and line graphs to illustrate findings.

# Exploratory Data Analysis (EDA) with SQL

**SQL Query Insights:**

- Identified unique launch sites and payload statistics.
- Analyzed successful and failed missions.
- Evaluated booster performance and payload capacities.

**Key Queries:**

- Query results provided insights into launch outcomes and booster performance across different time frames.

# Interactive Map - Folium

**Map Features:**

- **Launch Sites:** Visualized all launch sites on a global map.
- **Success Rates:** Color-coded markers indicating successful and failed launches.
- **Proximity Analysis:** Measured distances from launch sites to nearby infrastructure.

**Tools Used:**

- Folium for map creation and visualization.

# Dashboard with Plotly Dash

**Dashboard Features:**

- **Launch Success Visualization:** Displayed success rates across different sites.
- **Payload Analysis:** Analyzed the impact of payload mass on launch outcomes.
- **Interactive Elements:** Included dropdowns, sliders, and scatter plots for user interaction.

**Tools Used:**

- Plotly Dash for dashboard creation.

# RESULTS

# Results Summary

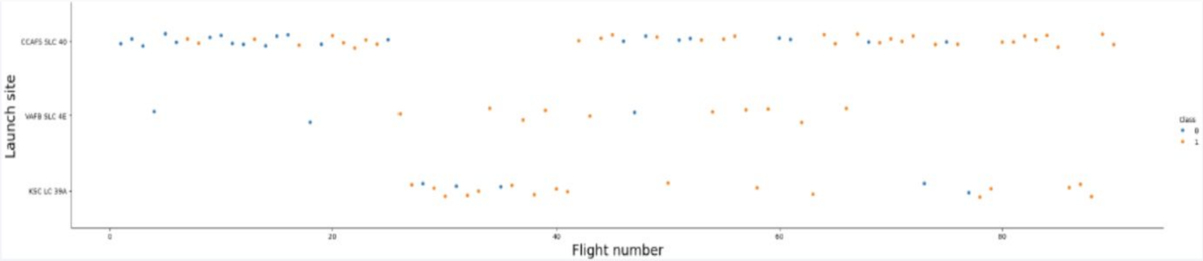
**Summary of Findings:**

- **EDA:** Revealed key insights into launch success factors.
- **Interactive Tools:** Demonstrated the power of dynamic visualization.
- **Predictive Modeling:** Identified the best-performing model for this dataset.

# Flight Number vs Launch Site

**Explanation:**

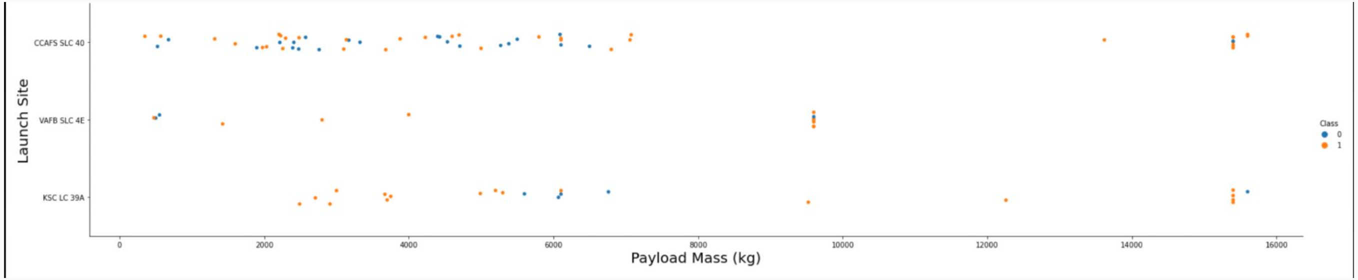
- **Initial Flights:** The first set of launches experienced failures, whereas the more recent ones were successful.
- **CCAFS SLC 40:** This launch site is responsible for nearly half of all SpaceX launches.
- **Success Rates:** The VAFB SLC 4E and KSC LC 39A sites exhibit higher success rates compared to others.
- **Trend Observation:** There appears to be a pattern where newer launches are increasingly successful.



# Payload vs Launch Site

**Explanation:**

- **Payload Mass and Success Rate:** Across all launch sites, an increase in payload mass generally correlates with a higher success rate.
- **Heavy Payloads:** The majority of launches carrying payloads exceeding 7000 kg achieved success.
- **KSC LC 39A Performance:** This site boasts a perfect success rate for launches with payloads under 5500 kg.



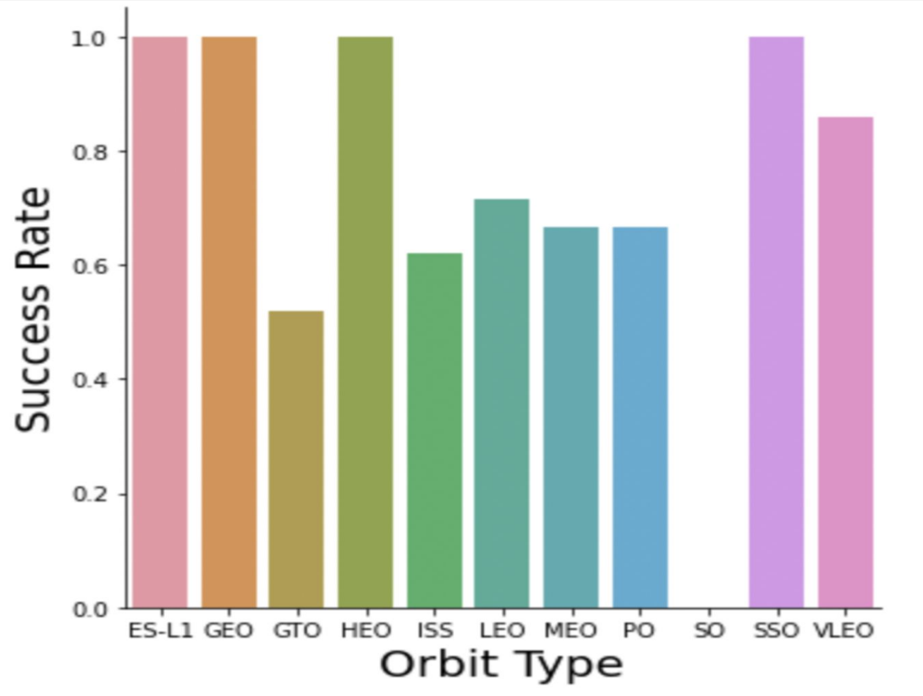
# Success Rate vs Orbit

**Modeling Approach:**

- **Data Preparation:** Standardized data and split into training/testing sets.
- **Model Evaluation:** Applied and compared multiple classification models (Logistic Regression, SVM, Decision Tree, KNN).
- **Results:** Decision Tree Model achieved the highest accuracy.

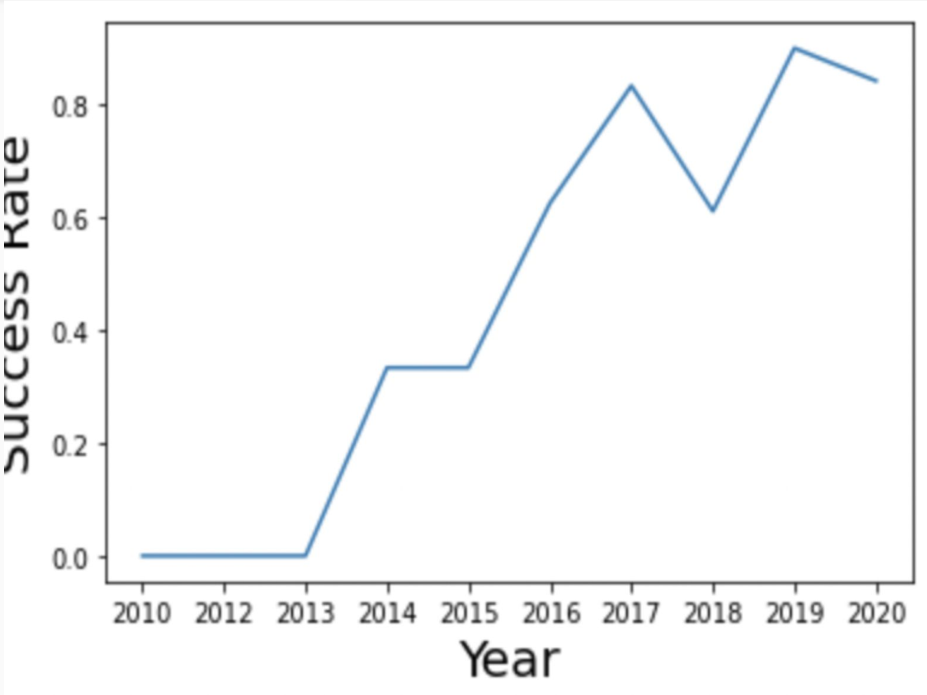
**Evaluation Metrics:**

- Confusion Matrix, Jaccard Score, and F1 Score to assess model performance.



# Launch Success

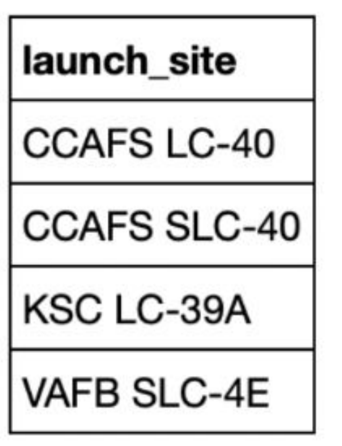
**Success Rate:** Overall success rate increased from 2013



# EDA WITH SQL

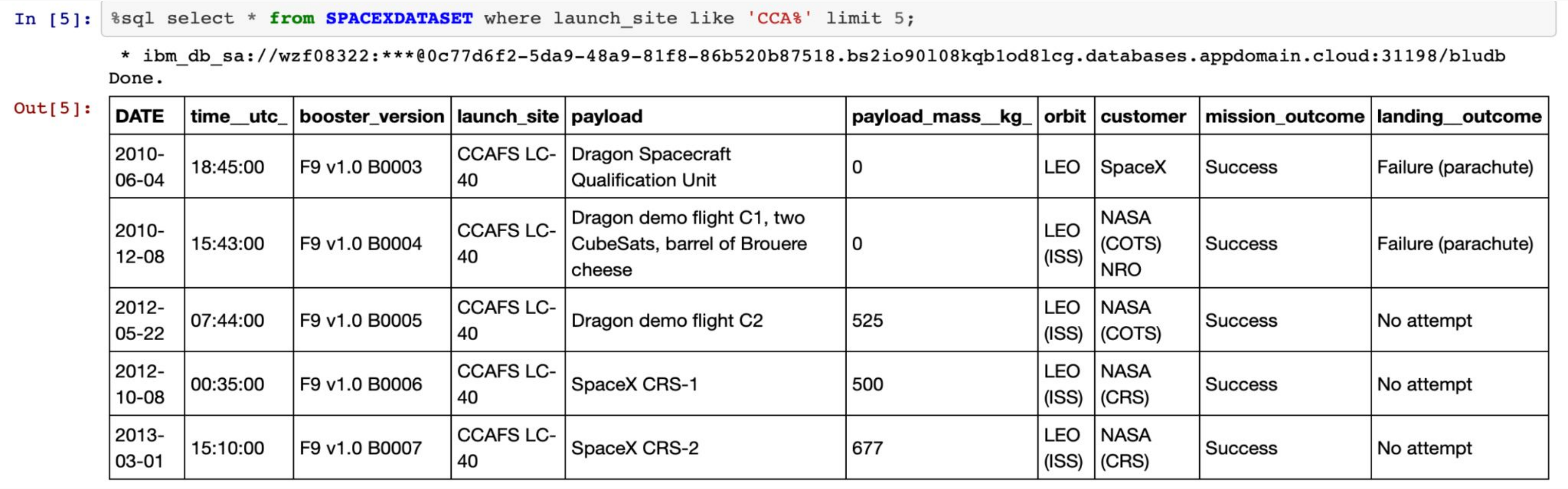
# Launch Site Names

**Explanation:** This shows all the unique launch site names



| launch_site |
|---|
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |

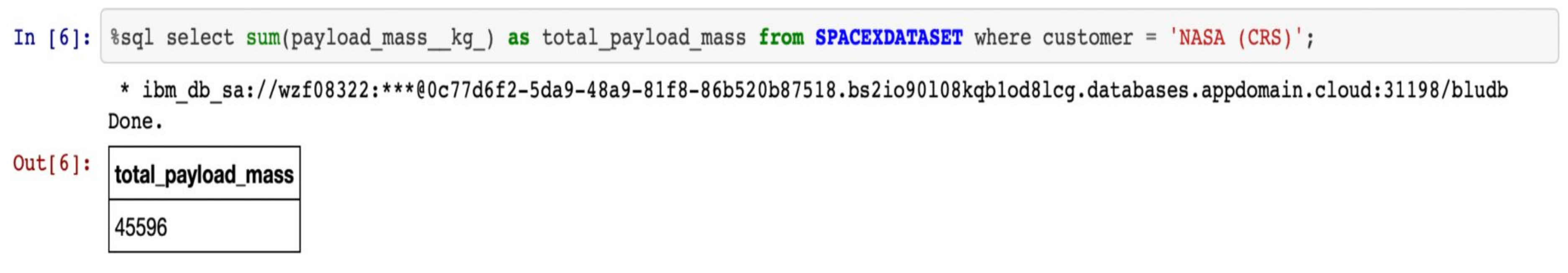
# Launch Site Names (Beginning with CCA)



```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[5]:
```

| 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 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 00:35:00 | 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



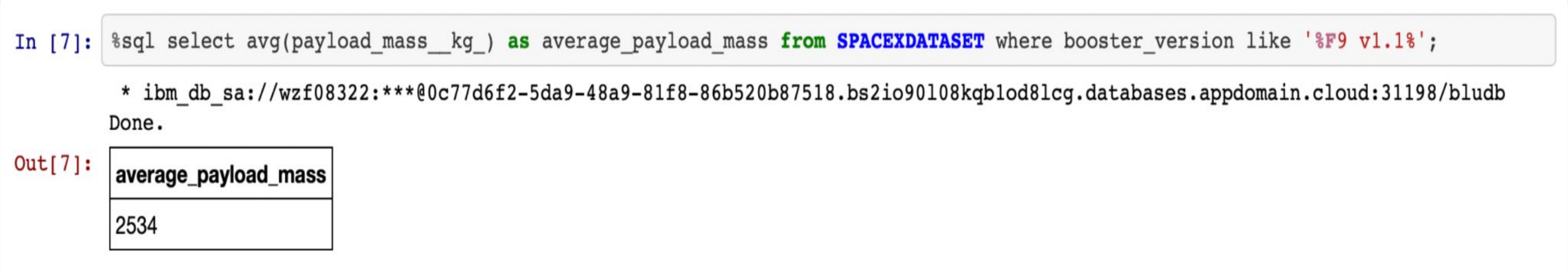
```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

| total_payload_mass |
|---|
| 45596 |

**Explanation:** total payload mass carried away by boosteraunched by NASA (CRS)

# Average Payload Mass

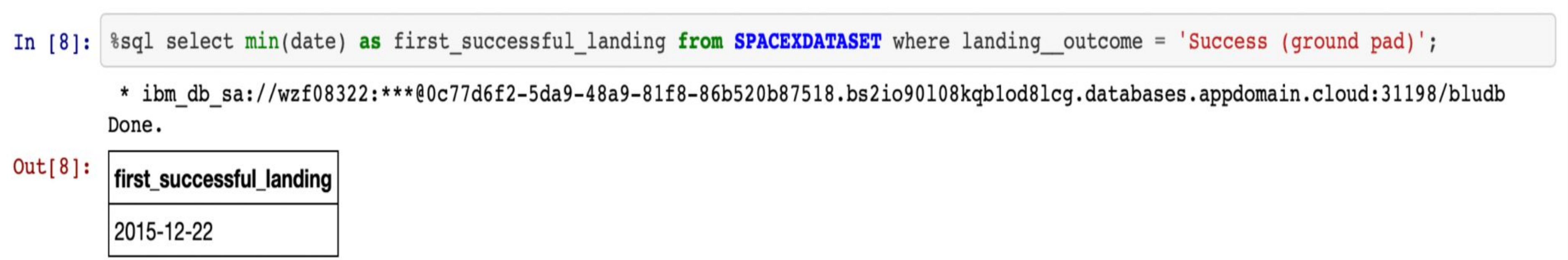


```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[7]:
```

| average_payload_mass |
|---|
| 2534 |

**Explanation:** Avg payload mass carried away by boosteraunched by NASA (F9 Vv1.1)

# First Successful Ground Landing



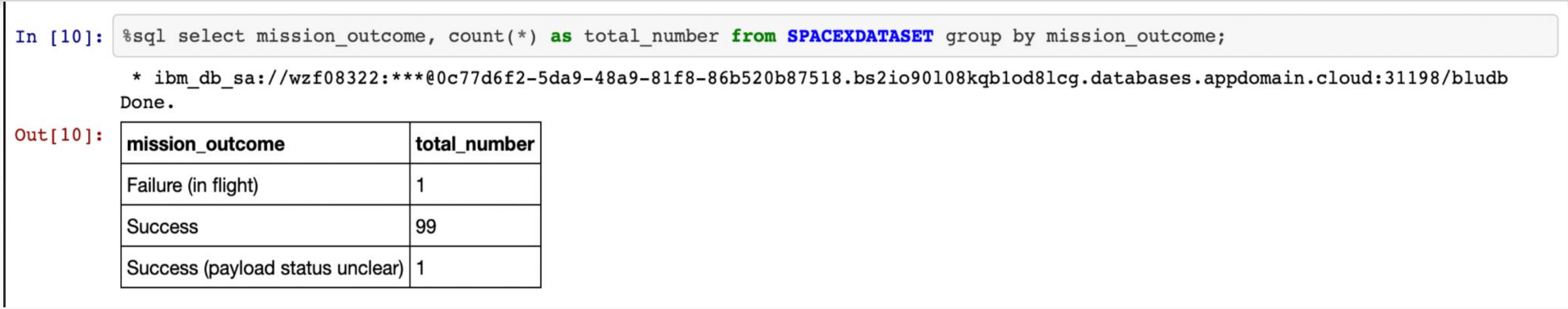
```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

| first_successful_landing |
|---|
| 2015-12-22 |

**Explanation:** first successful landing outcome in ground pad was achieved.

# Successful and Failed Missions

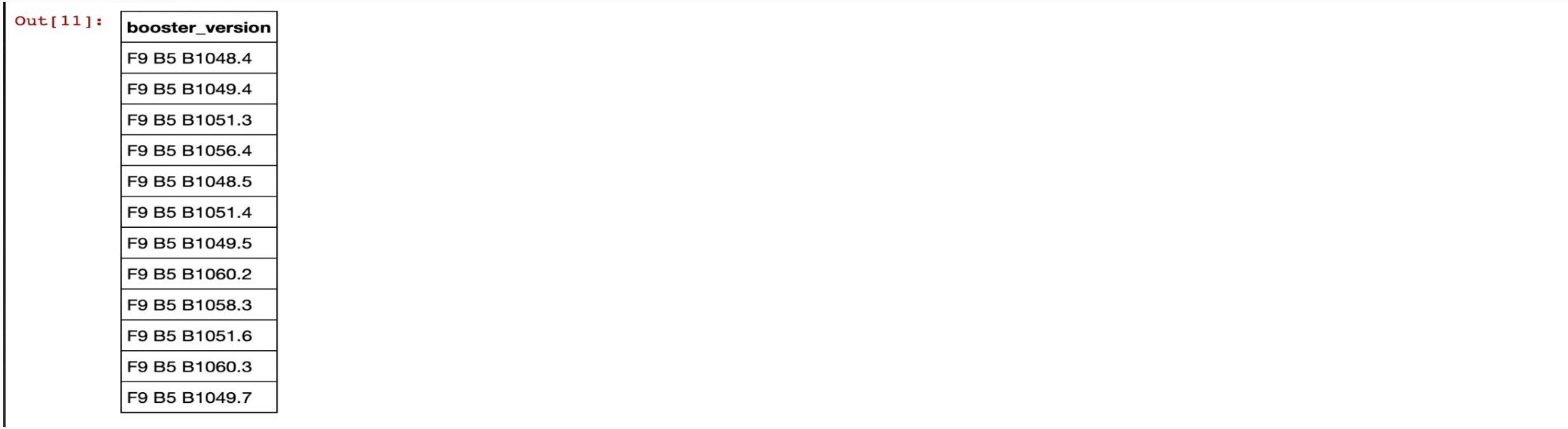


```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[10]:
```

| mission_outcome | total_number |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

**Explanation:** 1 Failure in Flight, 99 Success and 1 Success (payload status unclear)

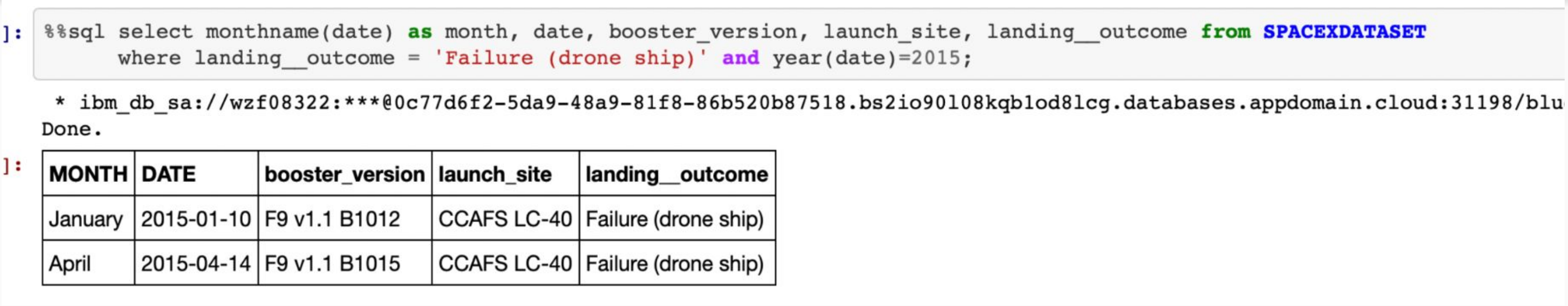
# Boosters (Carrying Max Payload)



Out[11]:

| booster_version |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1049.4 |
| F9 B5 B1051.3 |
| F9 B5 B1056.4 |
| F9 B5 B1048.5 |
| F9 B5 B1051.4 |
| F9 B5 B1049.5 |
| F9 B5 B1060.2 |
| F9 B5 B1058.3 |
| F9 B5 B1051.6 |
| F9 B5 B1060.3 |
| F9 B5 B1049.7 |

# Failed Landing on Drone ships (2015)



```
]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
         where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

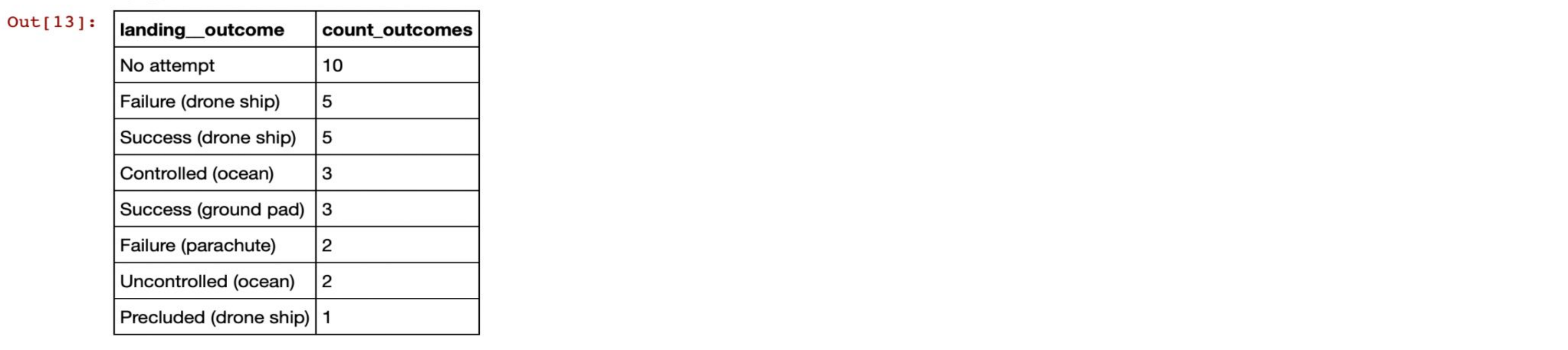
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/blu
Done.
```

]:

| MONTH | DATE | booster_version | launch_site | landing__outcome |
|---|---|---|---|---|
| January | 2015-01-10 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| April | 2015-04-14 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

**Explanation:** Showing month, date, booster version, launch site and landing outcome

# Successful Landing (Descending Order)



Out[13]:

| landing__outcome | count_outcomes |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

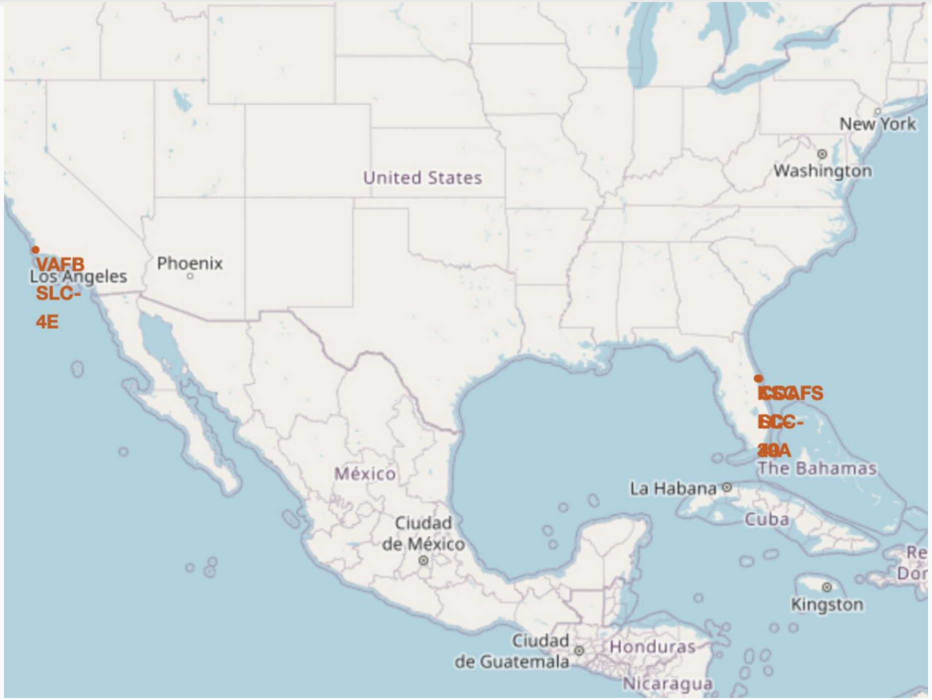
**Explanation:** Count of successful landing outcomes between 2010-06-04 and 2017-03-20

# INTERACTIVE MAP WITH FOLIUM

# Launch Sites

**Explanation:**

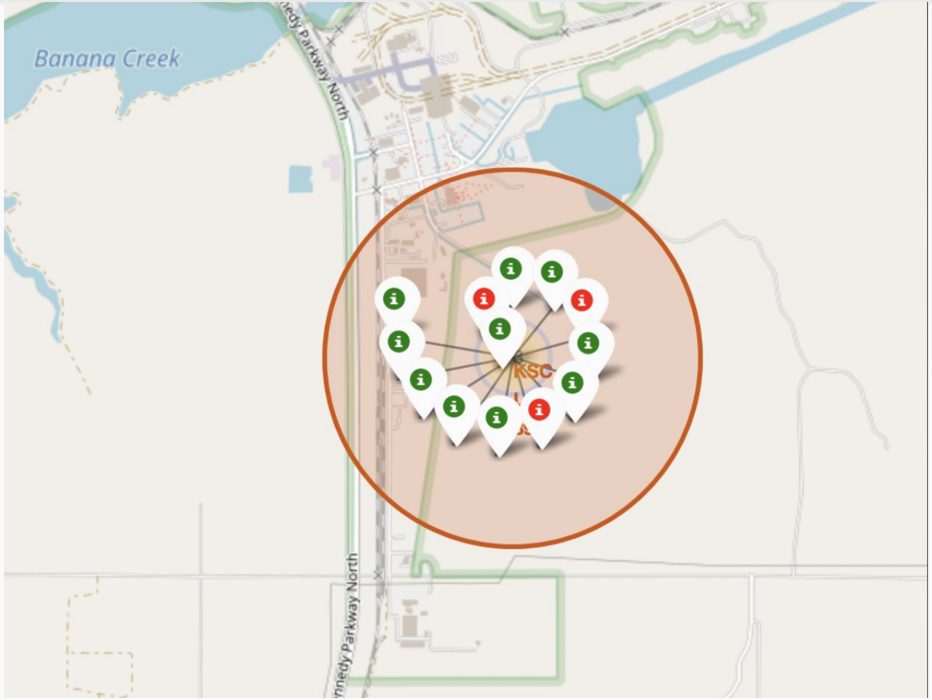
- **Proximity to the Equator:** The majority of launch sites are located near the Equator, where the Earth's rotation speed is fastest, approximately 1670 km/hour. This rotational velocity provides an additional boost to rockets launched from this region, aiding them in maintaining sufficient speed to achieve and sustain orbit due to inertia.
- **Coastal Locations:** All launch sites are situated close to the coast, which allows rockets to be launched over the ocean. This reduces the risk of debris falling or explosions occurring in populated areas, enhancing safety.



# Launch Outcomes

**Explanation:**

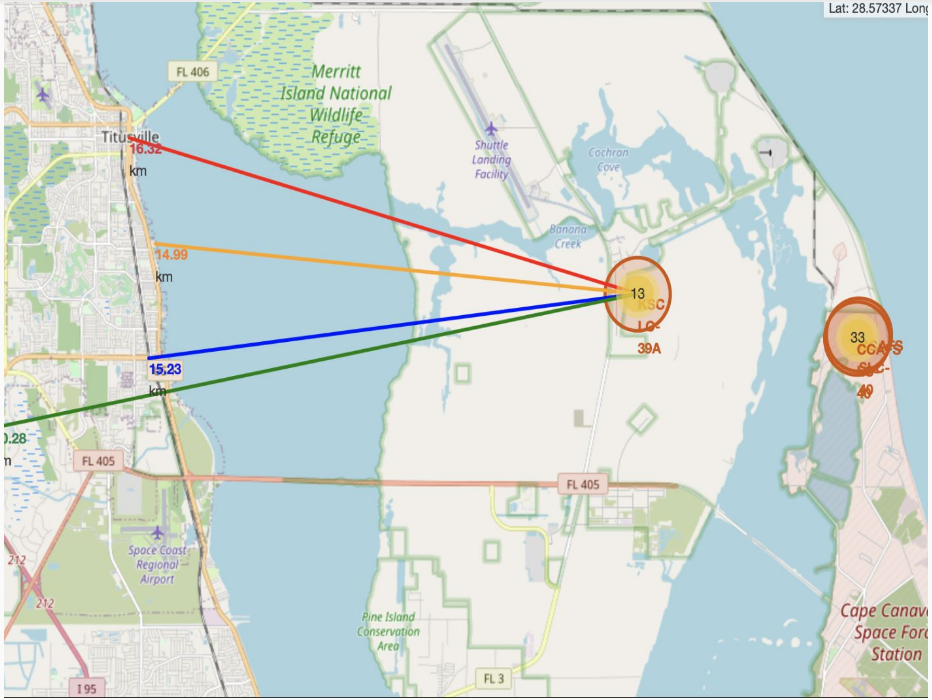
- **Color-Coded Markers:** The color-coded markers on the map allow for easy identification of launch sites with higher success rates.
    - **Green Marker:** Indicates a successful launch.
    - **Red Marker:** Indicates a failed launch.
- **KSC LC-39A Success:** The KSC LC-39A launch site demonstrates a particularly high success rate.



# Distance to Proximities

**Explanation:**

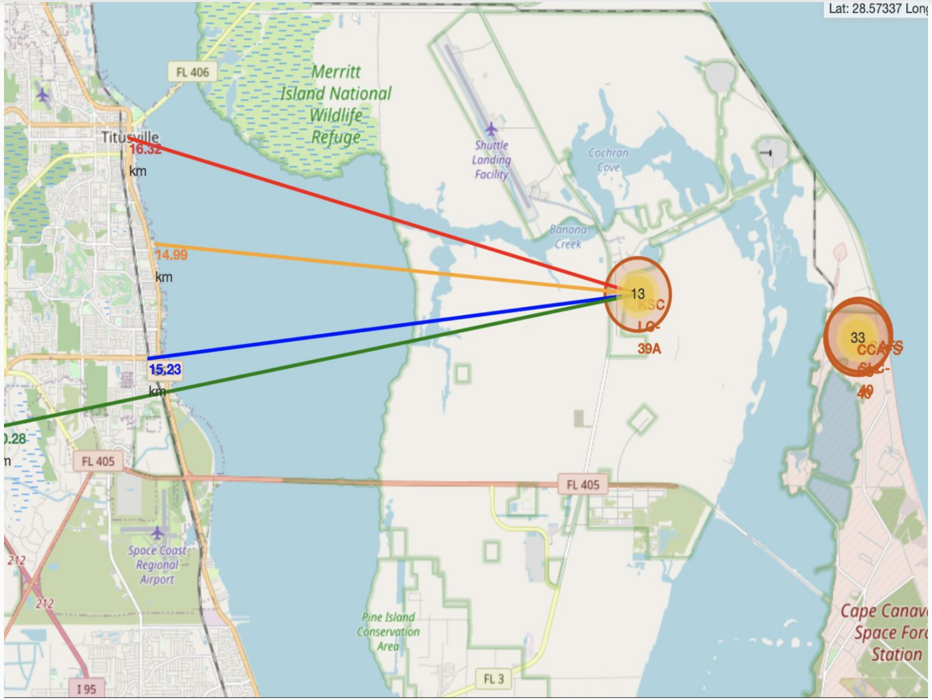
- **Proximity of KSC LC-39A:** The visual analysis reveals that the KSC LC-39A launch site is:
  - Approximately 15.23 km from the nearest railway.
  - Around 20.28 km from the closest highway.
  - About 14.99 km from the coastline.
- **Nearby City:** The launch site is also relatively close to Titusville, which is 16.32 km away.
- **Risk Consideration:** A failed rocket, traveling at high speed, could cover distances of 15-20 km within seconds, posing a potential risk to nearby populated areas.



# Distance to Proximities

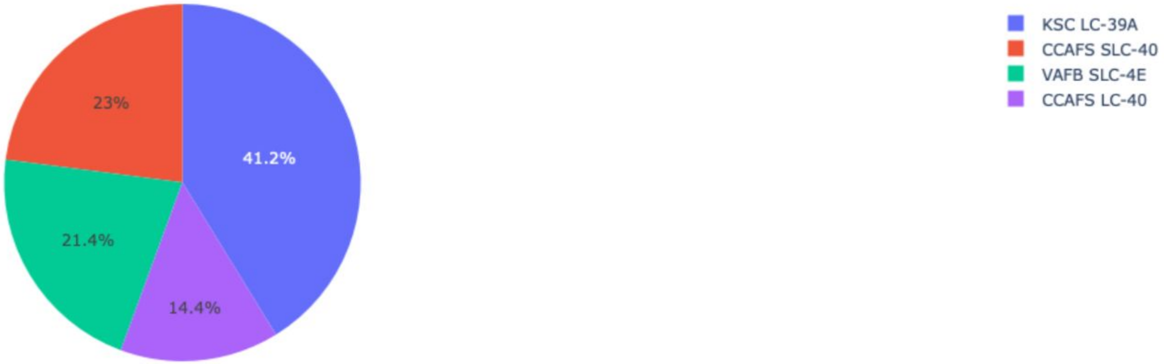
**Explanation:**

- **Proximity of KSC LC-39A:** The visual analysis reveals that the KSC LC-39A launch site is:
    - Approximately 15.23 km from the nearest railway.
    - Around 20.28 km from the closest highway.
    - About 14.99 km from the coastline.
- **Nearby City:** The launch site is also relatively close to Titusville, which is 16.32 km away.
- **Risk Consideration:** A failed rocket, traveling at high speed, could cover distances of 15-20 km within seconds, posing a potential risk to nearby populated areas.



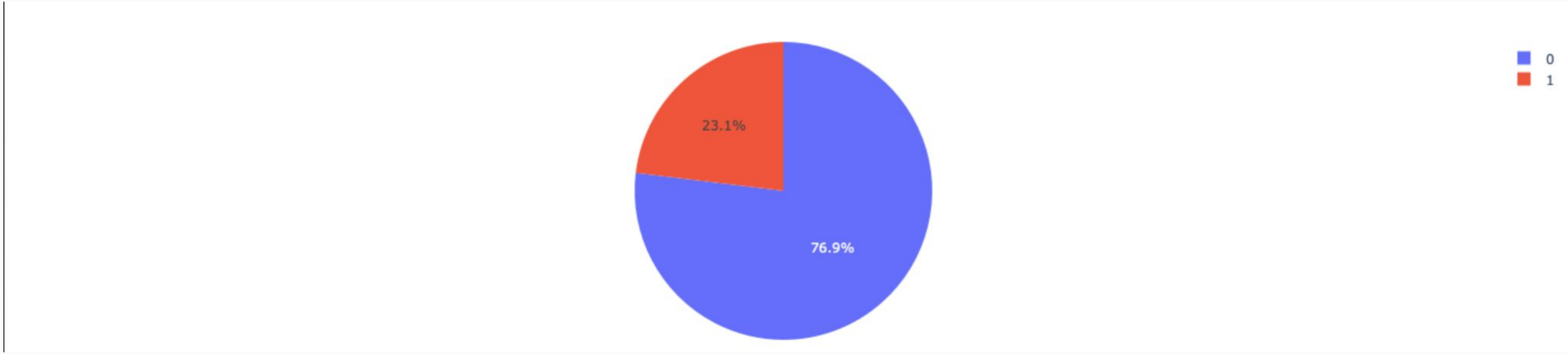
# DASHBOARD WITH PLOTLY

# Launch Success by Site



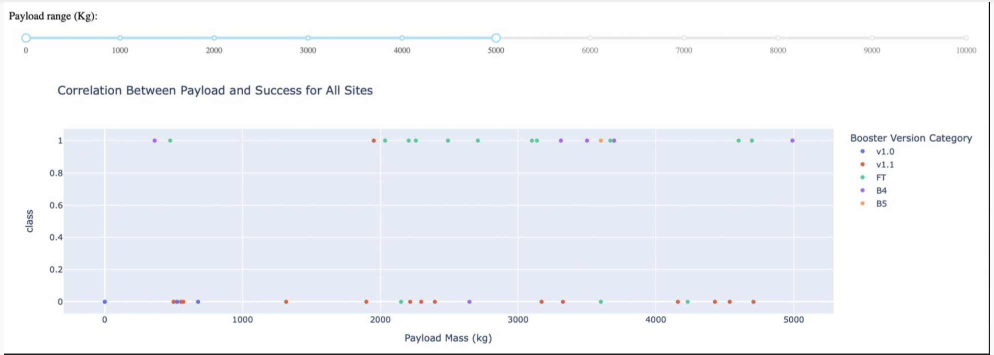
**Explanation:** The Graph shows KSC LC-39A has the most successful launches.

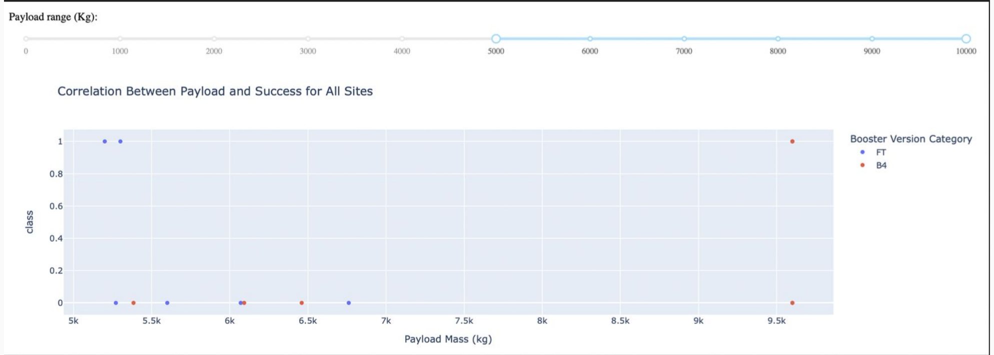
# Launch Success (KSC LC-29A)



**Explanation:** The Graph shows KSC LC-39A has the highest success rate amongst launch sites (76.9%)

# Payload Mass and Success (Booster Version)





**Explanation:** The Chart shows that the payload between 2000 and 5500 KG has the highest success rate

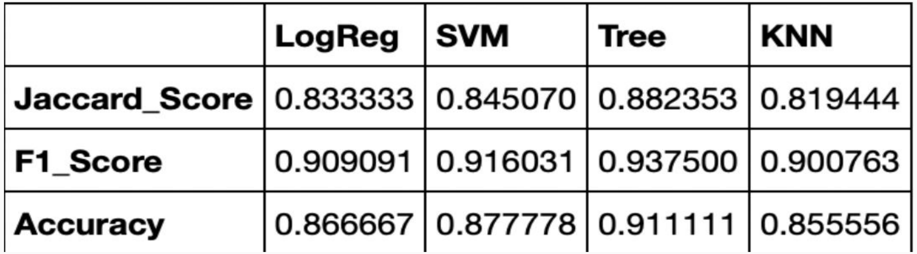
# PREDICTIVE ANALYSIS

# Classification (Accuracy)

**Explanation:**

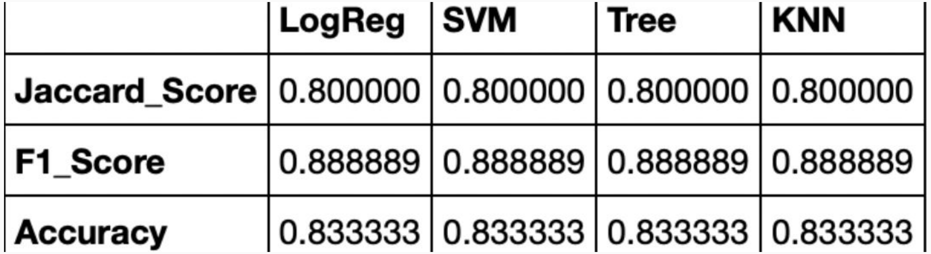
- **Test Set Results:** The test set scores do not definitively indicate which method performs best.
- **Sample Size Consideration:** The similar scores across methods may be attributed to the small test sample size of 18. Consequently, we extended the evaluation to the entire dataset.
- **Overall Dataset Analysis:** The analysis of the entire dataset reveals that the Decision Tree Model is the top performer, achieving not only the highest scores but also the greatest accuracy.

Test



| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.833333 | 0.845070 | 0.882353 | 0.819444 |
| F1_Score | 0.909091 | 0.916031 | 0.937500 | 0.900763 |
| Accuracy | 0.866667 | 0.877778 | 0.911111 | 0.855556 |

Entire Data Set

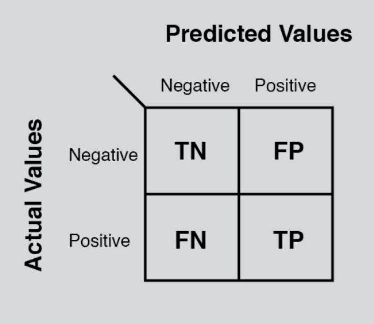


| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.800000 | 0.800000 | 0.800000 | 0.800000 |
| F1_Score | 0.888889 | 0.888889 | 0.888889 | 0.888889 |
| Accuracy | 0.833333 | 0.833333 | 0.833333 | 0.833333 |

# Confusion Matrix

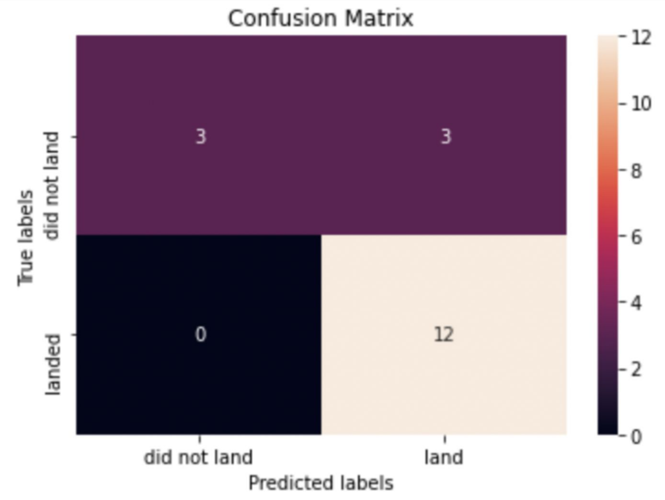
**Explanation:**

- **Confusion Matrix Analysis:** The confusion matrix indicates that logistic regression is capable of differentiating between the classes. However, the primary issue identified is the occurrence of false positives.



**Predicted Values**

| Actual Values | Negative | Positive |
|---|---|---|
| Negative | TN | FP |
| Positive | FN | TP |



# Conclusion

**Decision Tree Model:** The Decision Tree Model emerged as the most effective algorithm for this dataset.

**Payload Mass Impact:** Launches with lower payload masses tend to achieve higher success rates compared to those with heavier payloads.

**Launch Site Locations:** The majority of launch sites are situated near the Equator and are all closely located to the coastlines.

**Increasing Success Rate:** The success rate of launches has shown a consistent upward trend over the years.

**Top Performing Site:** Among all the launch sites, KSC LC-39A boasts the highest success rate.

**Orbit Success:** Orbits ES-L1, GEO, HEO, and SSO have maintained a perfect 100% success rate.

# Thanks!