IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Varshilkumar Patel
20th Nov 2024

# Outline

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

# Executive Summary

- **Summary of methodologies**
- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classiication)
- **Summary of all results**
- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

# Introduction

**Project Background and Context**

SpaceX has revolutionized the commercial space industry by making space travel more affordable. The company offers Falcon 9 rocket launches at $62 million per launch, significantly lower than the $165 million charged by other providers. This cost reduction is primarily due to SpaceX's ability to reuse the rocket's first stage. Predicting whether the first stage will successfully land is key to estimating the cost of a launch. Using publicly available data and machine learning models, this project aims to predict the reusability of the first stage.

**Questions to Address**

1. How do factors like payload mass, launch site, flight count, and orbit type influence the success of the first stage landing?
2. Has the success rate of first stage landings improved over time?
3. Which algorithm performs best for binary classification in this scenario?

Section 1

# Methodology

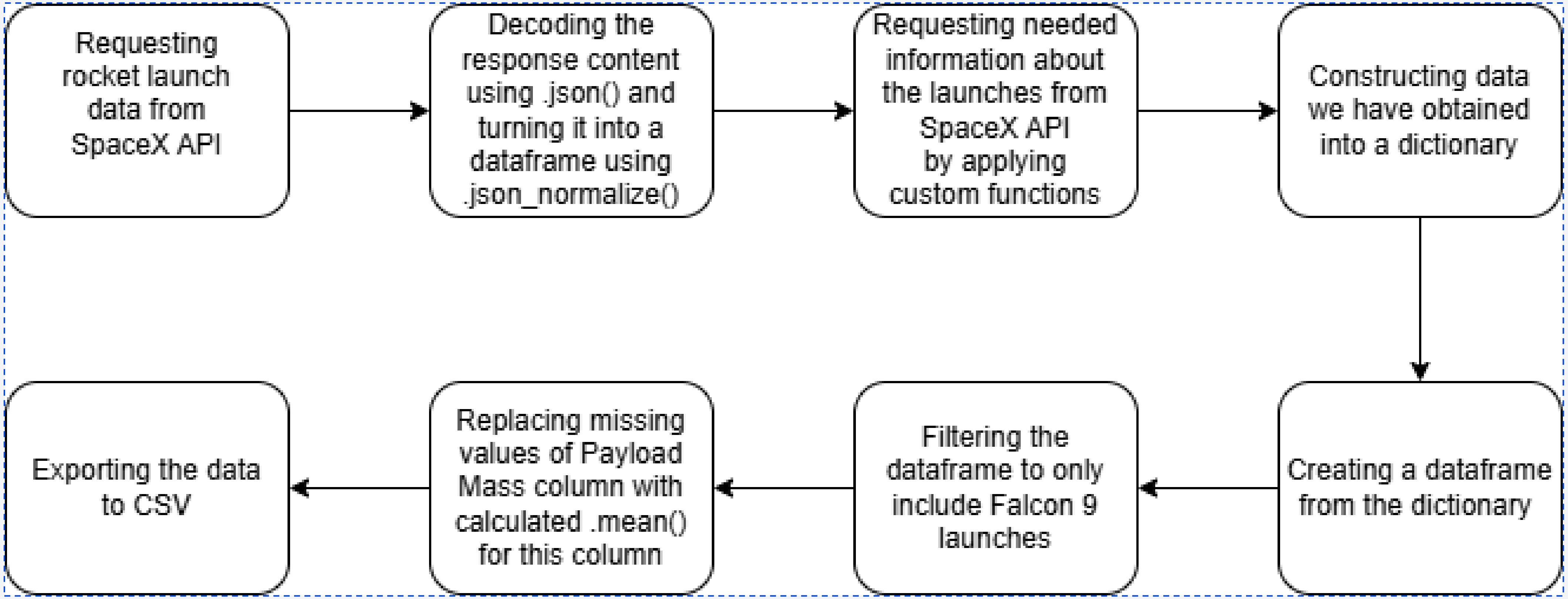
# Methodology

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classiication models to ensure the best results

# Data Collection

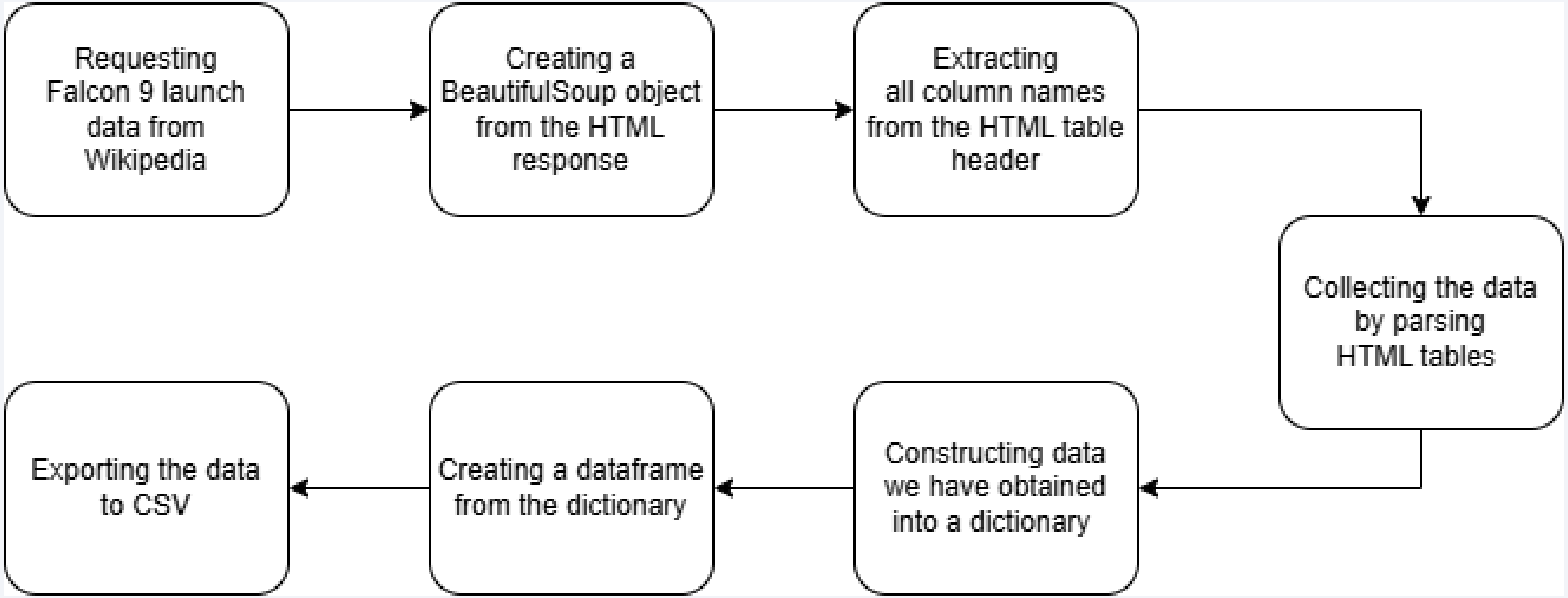
- The data collection process utilized a combination of API requests from the SpaceX REST API and web scraping from a table on SpaceX's Wikipedia page. Both methods were necessary to compile a complete dataset of launches for in-depth analysis.
- Data columns obtained through the SpaceX REST API include:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- Data columns obtained via Wikipedia web scraping include:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

# Data Collection – SpaceX API

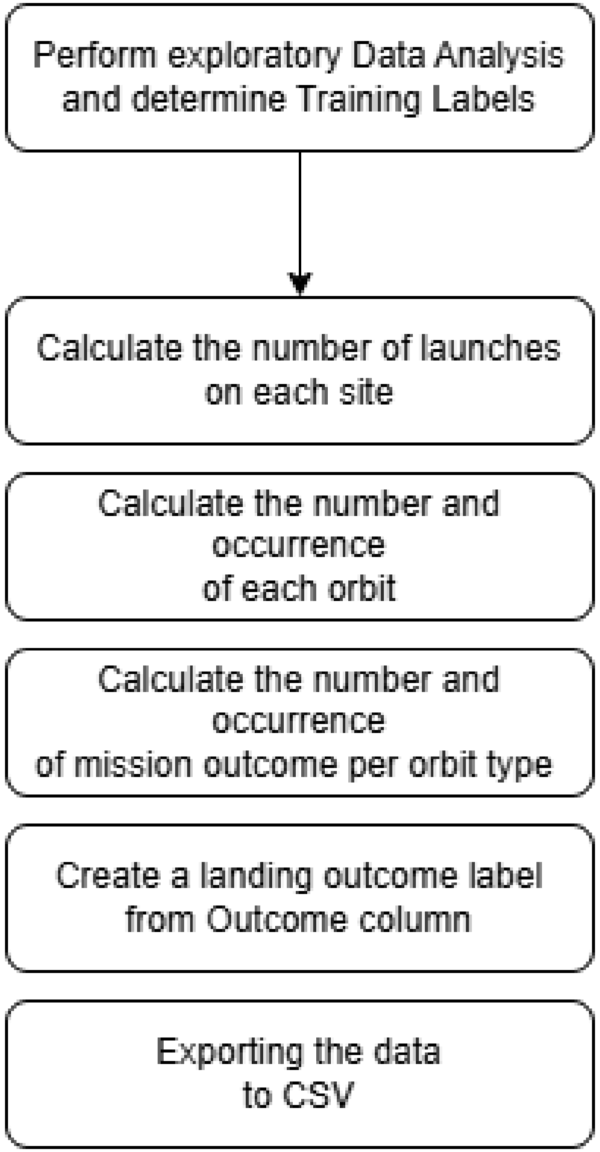



[Github URL: Data Collection SpaceX API]

# Data Collection - Scraping




[Github URL: Web Scraping]




# Data Wrangling

- The dataset includes various instances where the booster did not land successfully. In some cases, a landing was attempted but failed due to an accident. For example, "True Ocean" indicates the mission successfully landed in a specific ocean region, whereas "False Ocean" signifies an unsuccessful landing in the same region. Similarly, "True RTLS" represents a successful landing on a ground pad, while "False RTLS" indicates failure. "True ASDS" denotes a successful landing on a drone ship, whereas "False ASDS" reflects an unsuccessful attempt. These outcomes are primarily converted into training labels, where "1" indicates a successful landing and "0" denotes a failure.

[Github URL: Data wrangling]

# EDA with Data Visualization

- The following charts were created:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Orbit Type vs. Success Rate
  - Flight Number vs. Orbit Type
  - Payload Mass vs. Orbit Type
- Success Rate Yearly TrendScatter plots illustrate the relationships between variables and can serve as inputs for machine learning models if a clear pattern is present.Bar charts compare discrete categories, highlighting the relationships between specific groups and their corresponding measured values.Line charts depict data trends over time, making them ideal for time series analysis.



[Github URL: EDA with Data Visualization]

# EDA with SQL

- Performed SQL queries:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the irst successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking 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

[Github URL: EDA with SQL]()

# Build an Interactive Map with Folium

- Markers of all Launch Sites:
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Coloured Markers of the launch outcomes for each Launch Site:
  - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
  - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

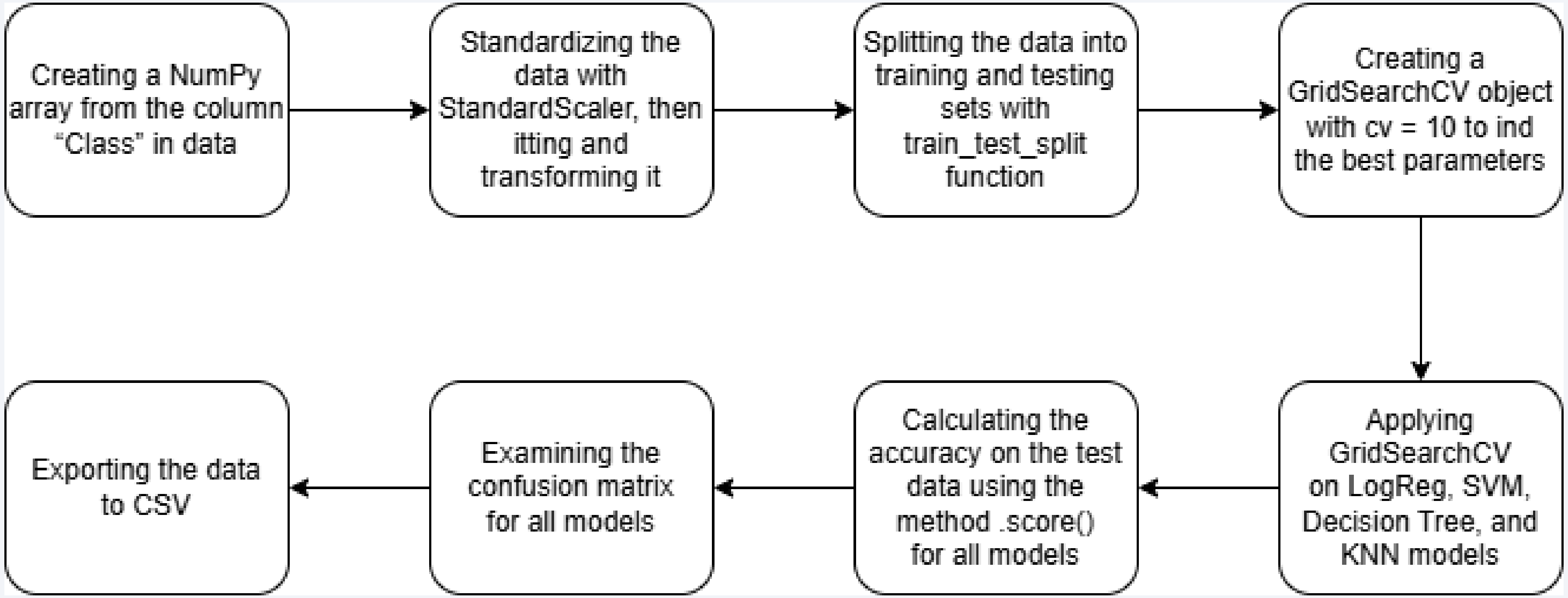
[Github URL: Interactive Visual Analytics with Folium]

# Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a speciic Launch Site was selected.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs. Success Rate for the dierent Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success.

[Github URL: SpaceX Dash App]

# Predictive Analysis (Classification)




[Github URL: Machine Learning Prediction]

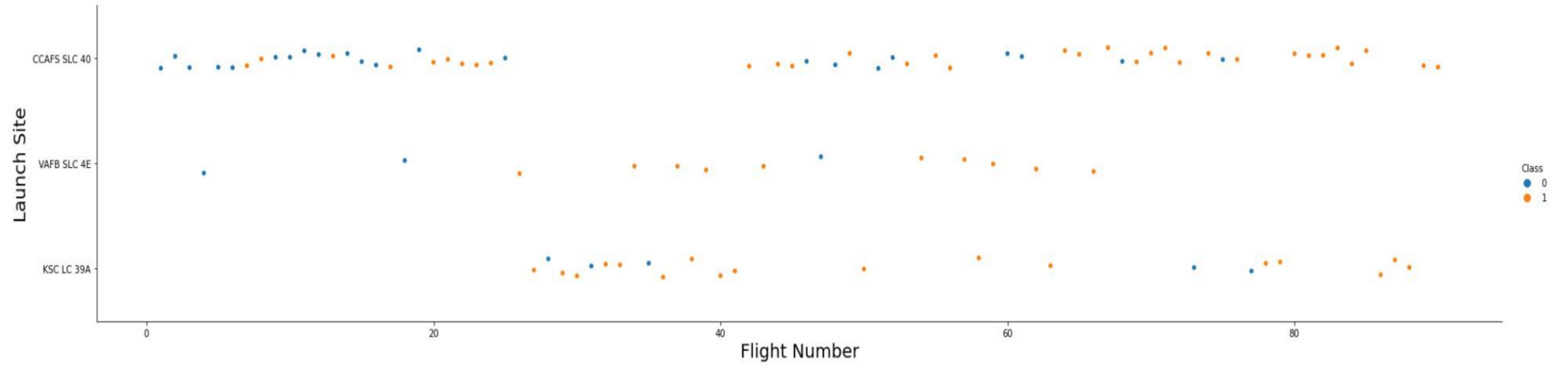
# Results

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

Section 2

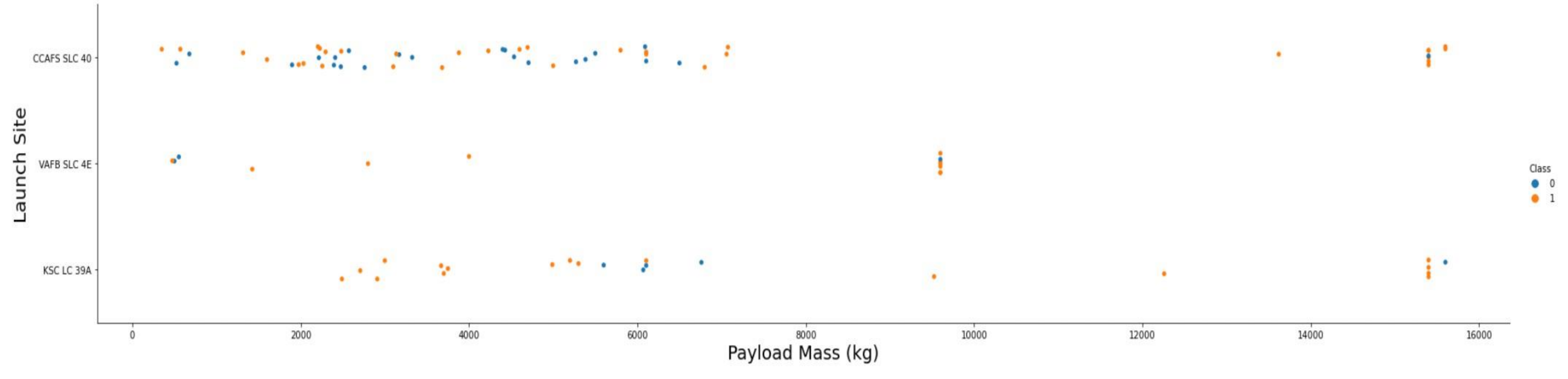
# Insights drawn from EDA

# Flight Number vs. Launch Site




- The initial flights experienced complete failure, whereas the most recent ones have achieved consistent success.
- The CCAFS SLC 40 launch site accounts for nearly half of all launches.
- The VAFB SLC 4E and KSC LC 39A sites demonstrate higher success rates.
- It can be inferred that the likelihood of success improves with each subsequent launch.

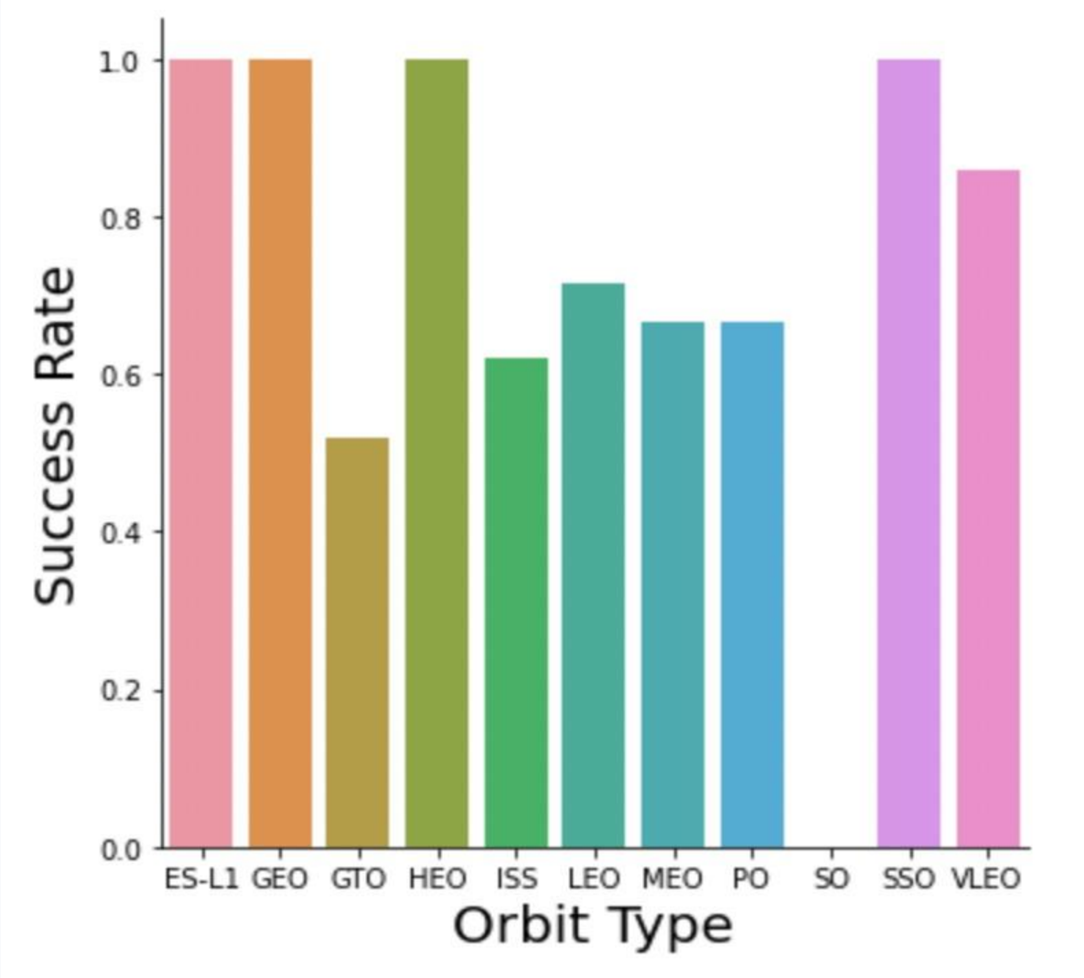
# Payload vs. Launch Site



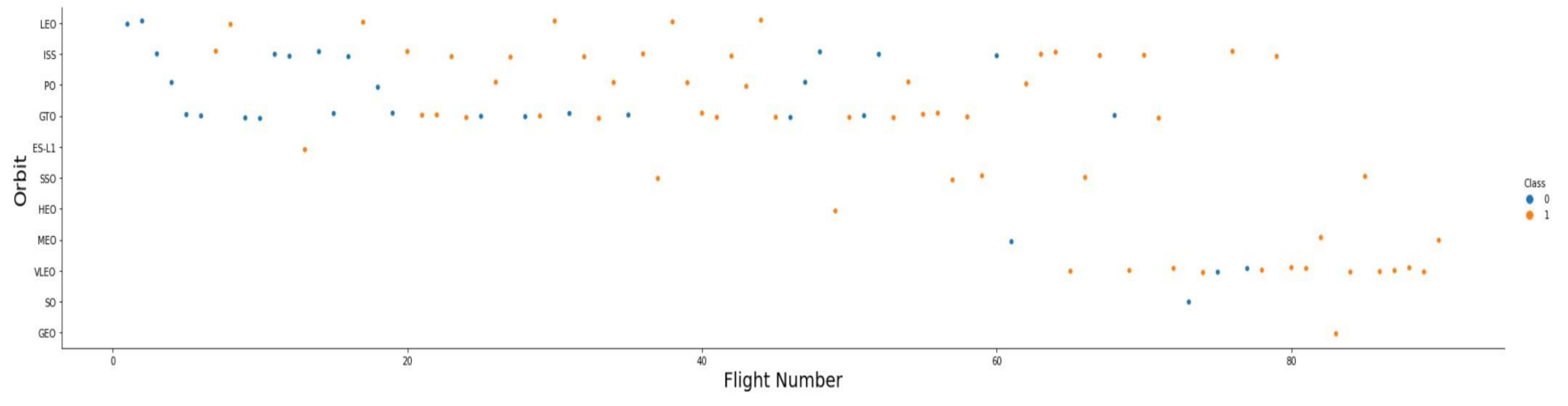

- Success rate increases with payload mass across all launch sites.
- The majority of launches with a payload mass exceeding 7000 kg were successful.
- Additionally, the KSC LC 39A site achieved a 100% success rate for payloads weighing less than 5500 kg.

# Success Rate vs. Orbit Type

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO

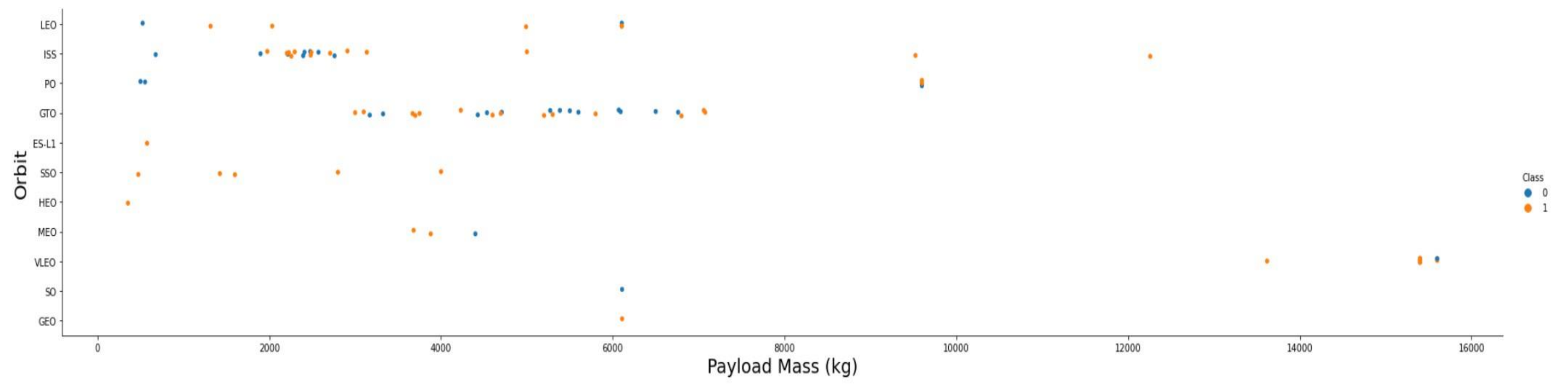


# Flight Number vs. Orbit Type




- In the LEO orbit, success seems to be influenced by the number of flights. However, in the GTO orbit, there appears to be no correlation between success and flight number.

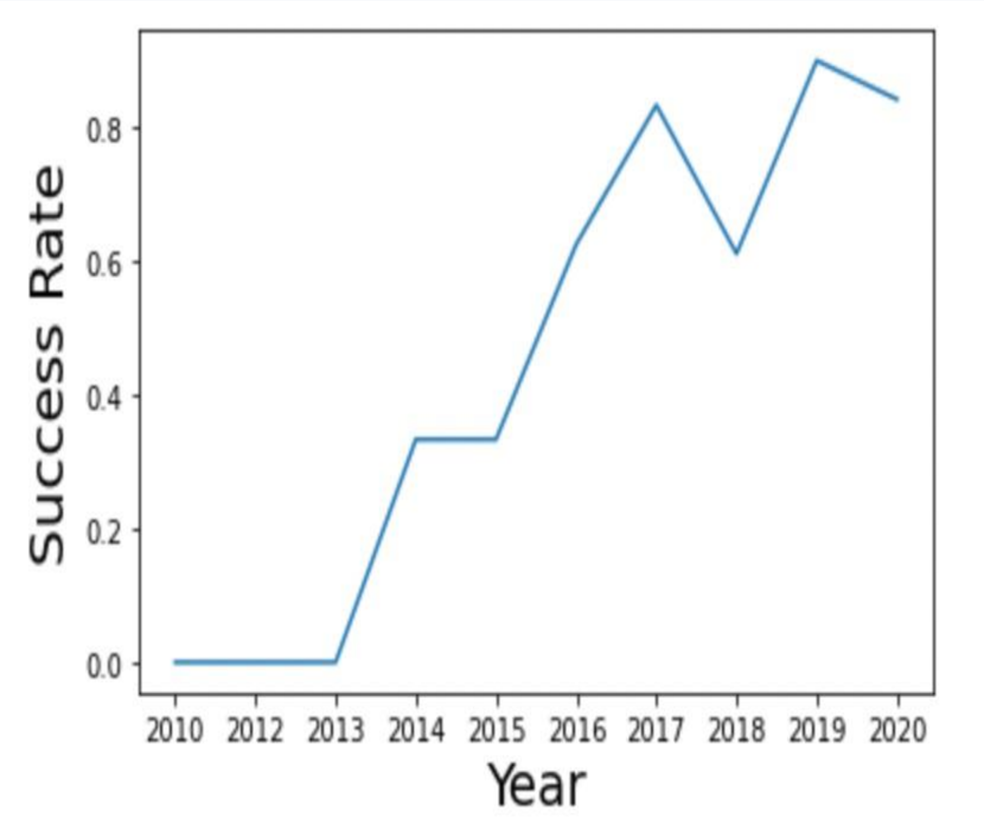
# Payload vs. Orbit Type




- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

# Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



# All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

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

Out[4]:
```

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

- Displaying the names of the unique launch sites in the space mission

# Launch Site Names Begin 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 |

- Displaying 5 records where launch sites begin with the string 'CCA'.

# 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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

| total_payload_mass |
|---|
| 45596 |

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

# Average Payload Mass by F9 v1.1

```
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 |

- Displaying average payload mass carried by booster version F9 v1.1.

# First Successful Ground Landing Date

```
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 |

- Listing the date when the first successful landing outcome in ground pad was achieved.

# Successful Drone Ship Landing with Payload between 4000 and 6000

In [9]:
```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;
```

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

Out[9]:

| booster_version |
|---|
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

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

# Total Number of Successful and Failure Mission Outcomes

```
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 |

- Listing the total number of successful and failure mission outcomes.

# Boosters Carried Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
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 |

- Listing the names of the booster versions which have carried the maximum payload mass.

# 2015 Launch Records

In [12]:
```
%%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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

| 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) |

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

In [13]:
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

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

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 |

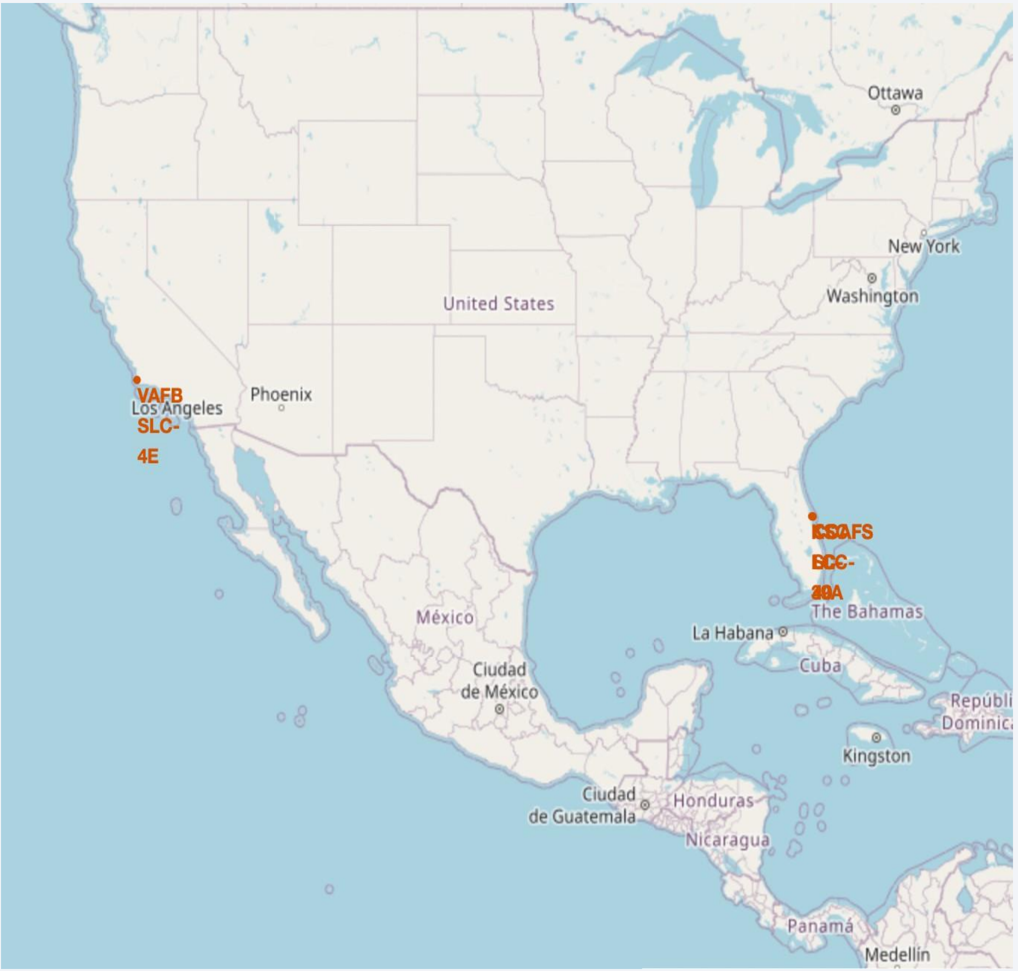
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

Section 3

# Launch Sites Proximities Analysis

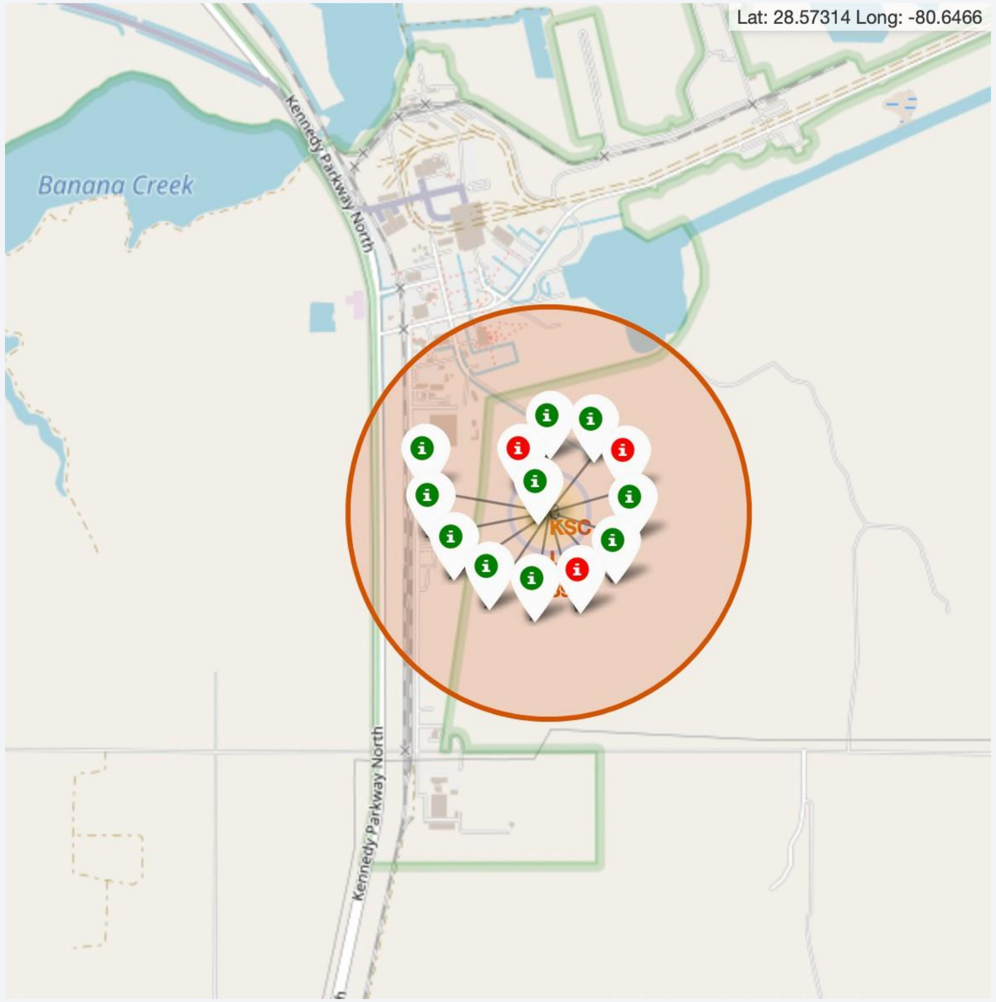
# All Launch Sites Location markers on global map

- Most launch sites are located near the equator, where the Earth's surface moves faster than at other latitudes. At the equator, the surface speed is approximately 1670 km/hour, and this velocity is carried into space by rockets due to inertia. This additional speed helps spacecraft achieve the velocity needed to stay in orbit more efficiently.
- Additionally, launch sites are typically situated close to the coast to reduce risks. By launching rockets over the ocean, the potential for debris or explosions causing harm to populated areas is minimized.



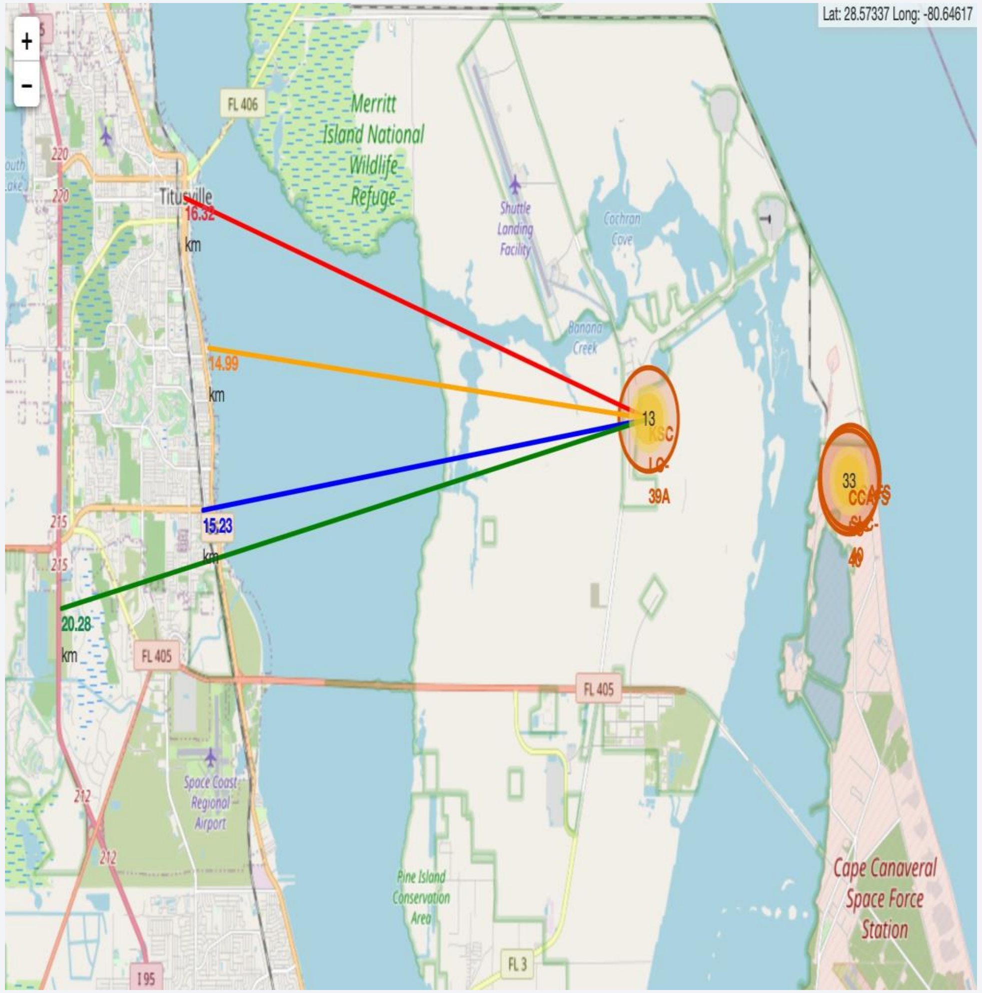
# Colors-labeled launch record on the map

- The color-coded markers make it easy to identify launch sites with relatively high success rates:
  - Green Marker: Successful Launch
  - Red Marker: Failed Launch
- The KSC LC-39A launch site stands out for its exceptionally high success rate.



# Distance from Launch Site KSC LC-39A to its proximities

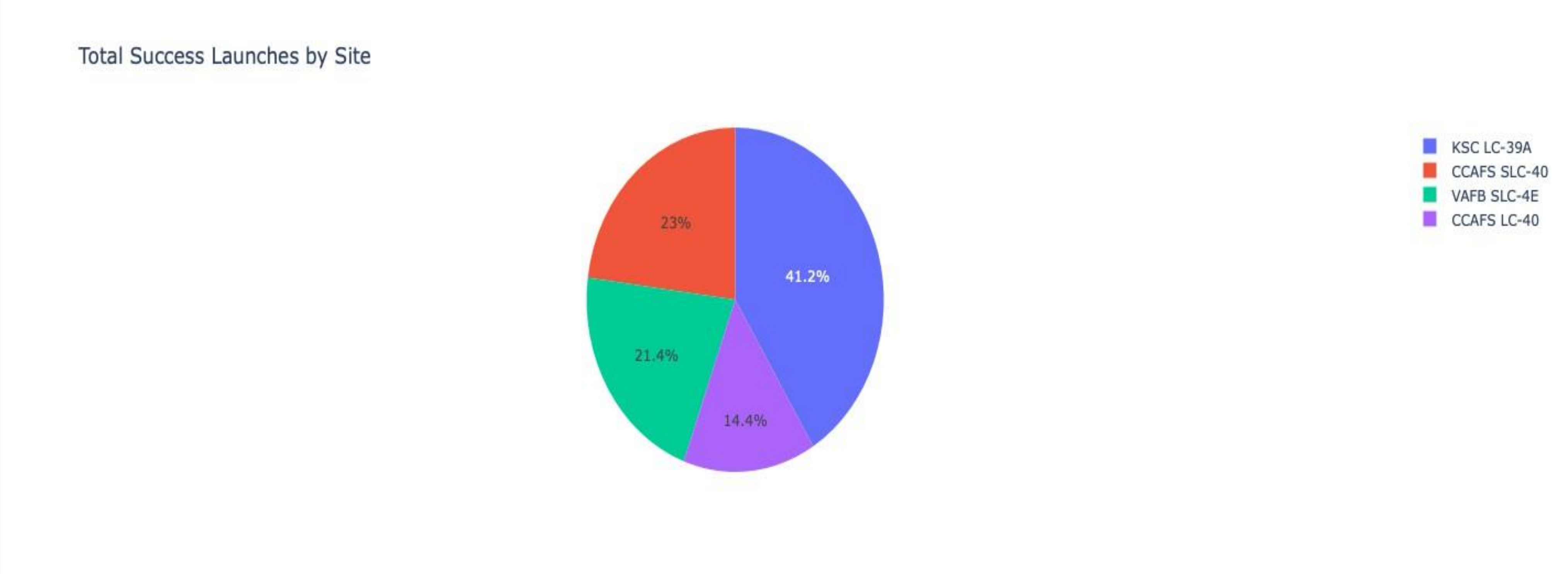
- he visual analysis of the KSC LC-39A launch site reveals its proximity to key locations:
- Railway: approximately 15.23 km
- Highway: approximately 20.28 km
- Coastline: approximately 14.99 km
- Closest city, Titusville: approximately 16.32 km
- Given the high speeds of failed rockets, they could cover distances of 15–20 km within seconds, posing potential risks to nearby populated areas.



Section 4

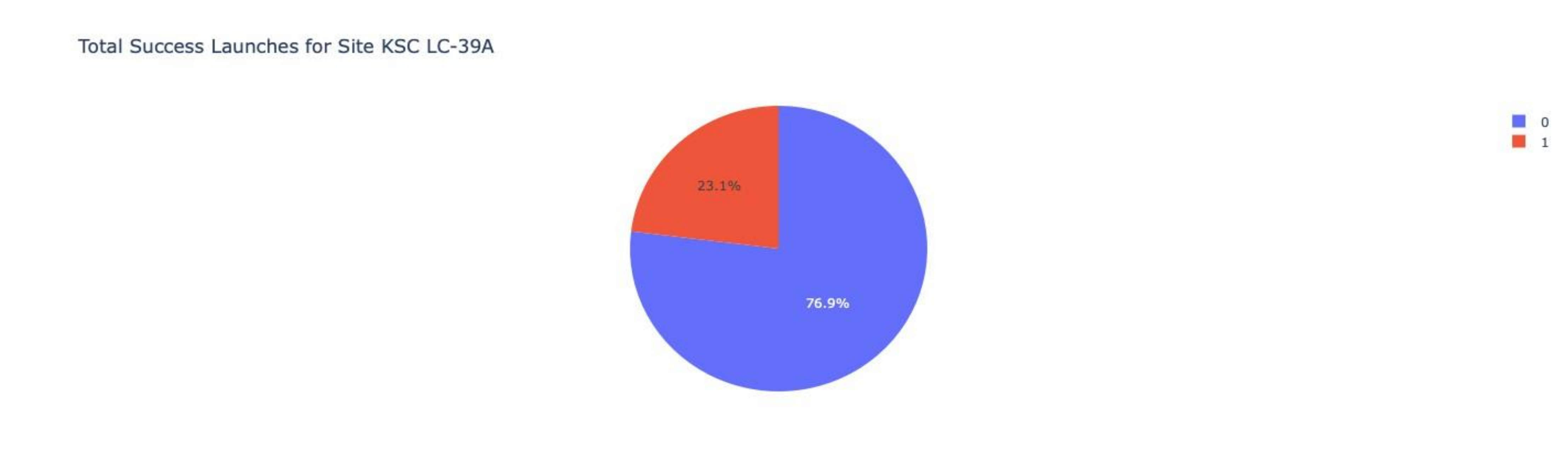
# Build a Dashboard with Plotly Dash

# Launch Success Count for all sites




- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

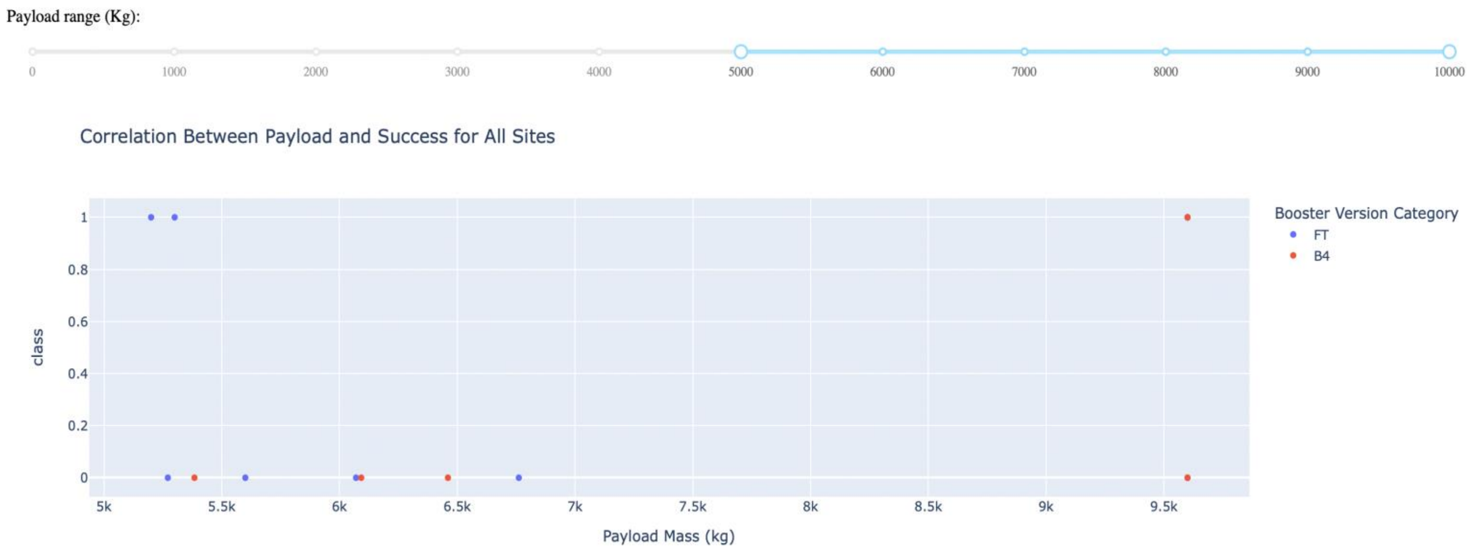
# Launch Site with highest launch success ratio




- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

# Payload Mass Vs. Launch outcome for all sites

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- The Test Set scores do not provide a clear indication of which method performs best, likely due to the small test sample size (18 samples). To address this, we evaluated all methods using the entire dataset. The results confirm that the Decision Tree Model is the most effective, achieving both higher overall scores and the highest accuracy.

Scores and Accuracy of the Test Set

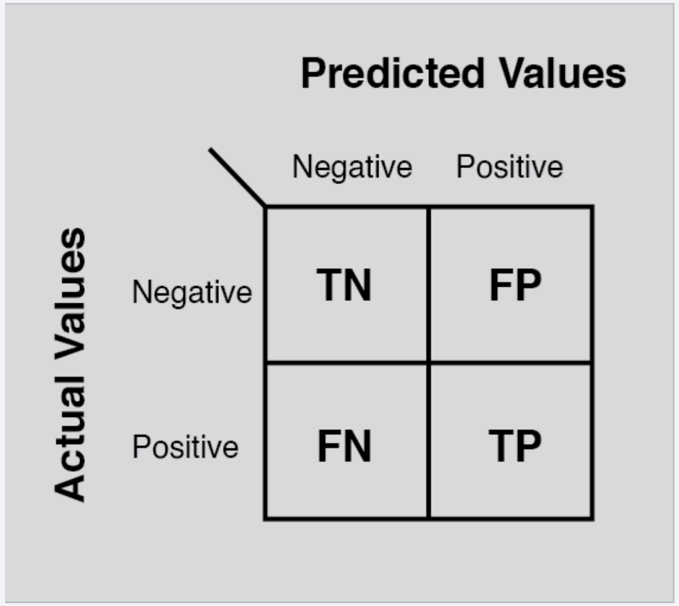
| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.800000 | 0.800000 | 0.800000 | 0.800000 |
| F1_Score | 0.888889 | 0.888889 | 0.888889 | 0.888889 |
| Accuracy | 0.833333 | 0.833333 | 0.833333 | 0.833333 |

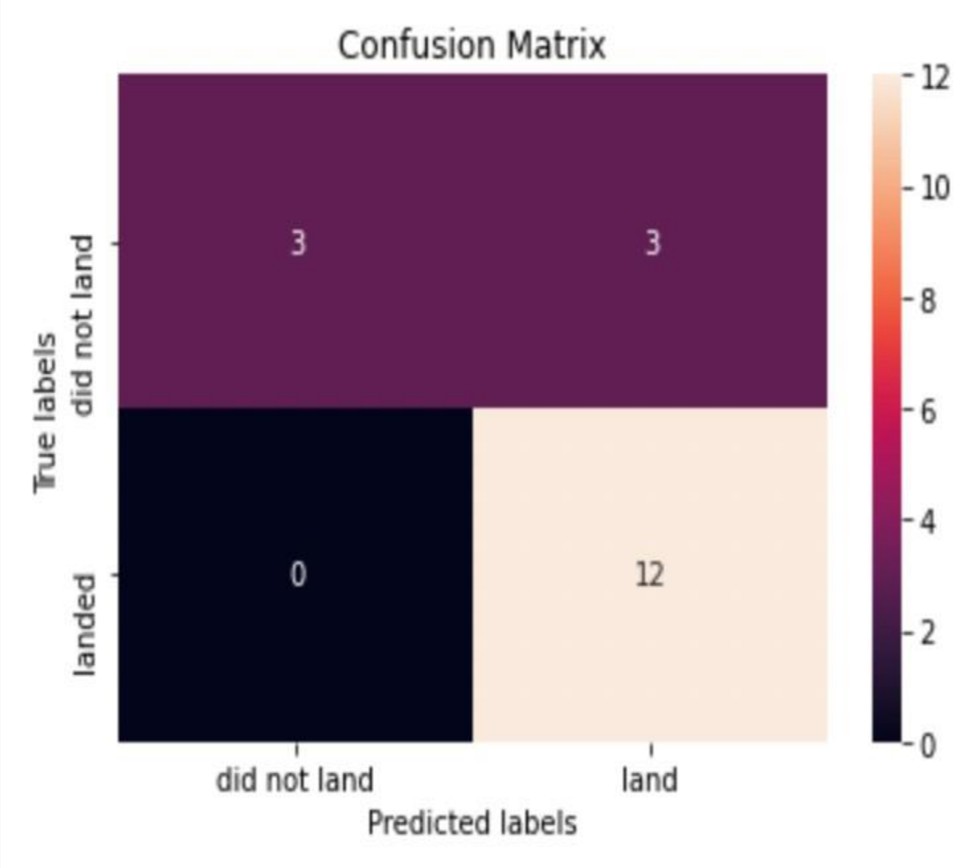
Scores and Accuracy of the Test Set

| | 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 |

# Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



# Conclusions

- The Decision Tree model proved to be the most effective algorithm for this dataset.
- Launches with smaller payload masses tend to yield better outcomes compared to those with heavier payloads.
- Most launch sites are located near the Equator and are situated very close to the coast.
- The success rate of launches has shown an upward trend over the years. Among all sites, KSC LC-39A boasts the highest success rate for launches.
- Orbits such as ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.

# Appendix

Special Thanks to:

Instructors

Coursera

IBM

# Thank you!