IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

David Reynolds
February 14, 2025

# 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 (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

# Introduction

- Project background and context
    - SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
    - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
    - Does the rate of successful landings increase over the years?
    - What is the best algorithm that can be used for binary classification in this case?

Section 1

# Methodology

# Methodology

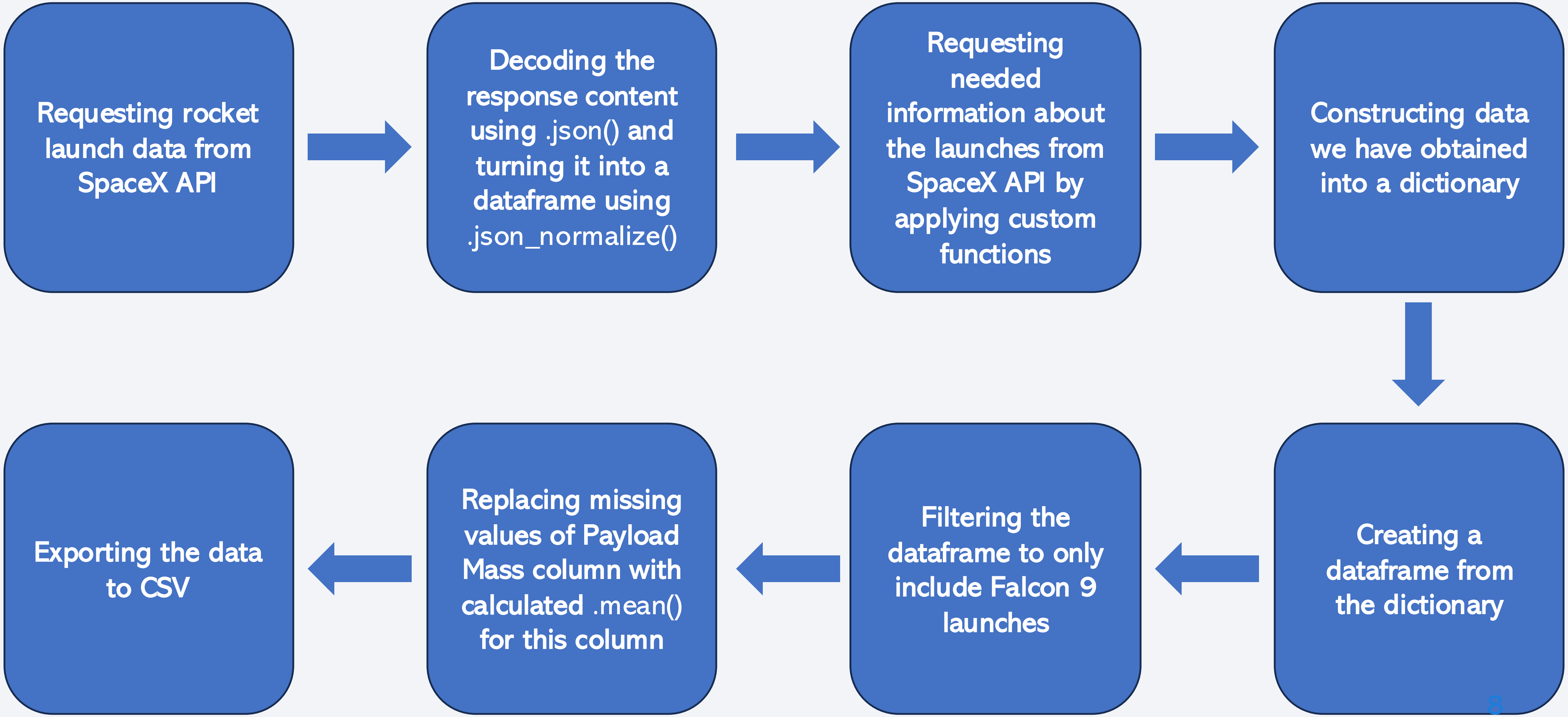
## Executive Summary

- 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

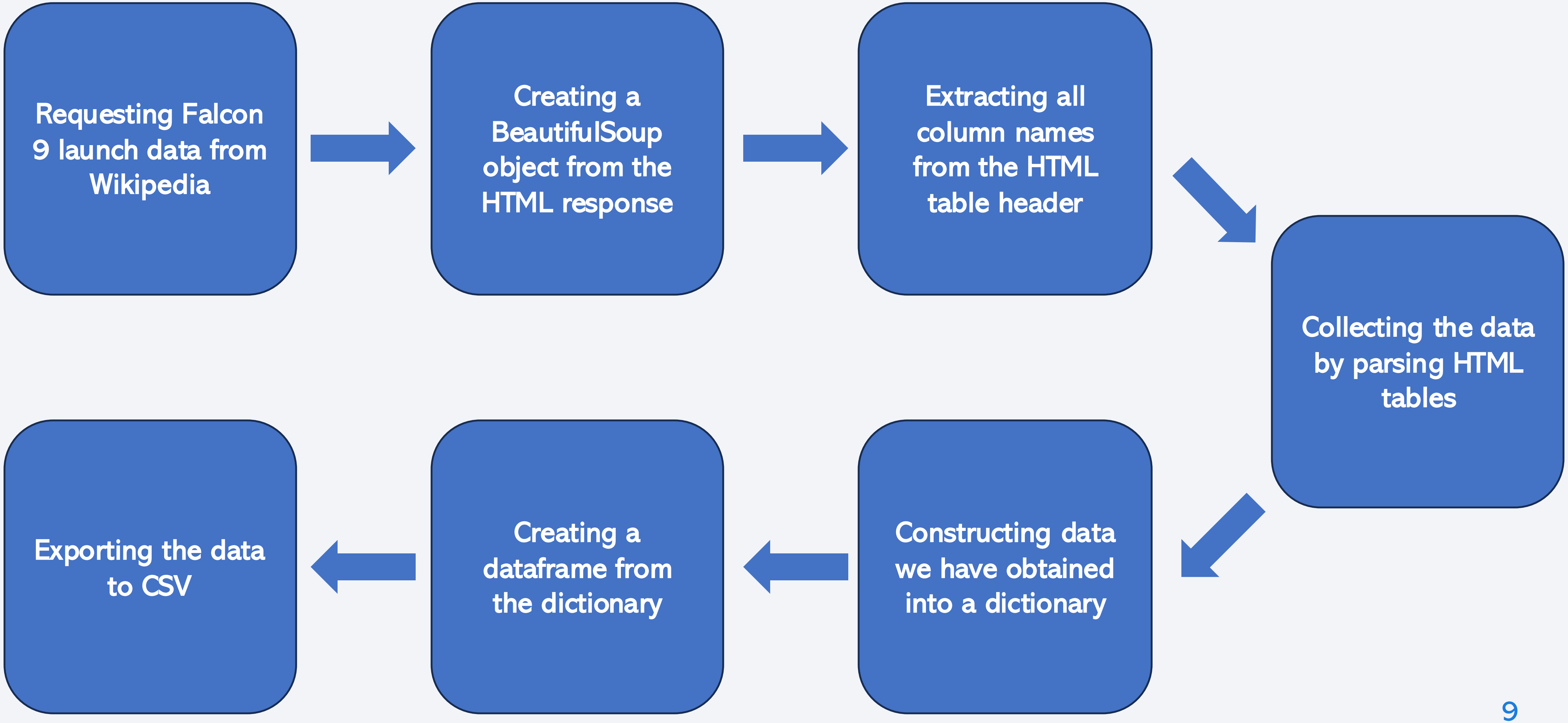
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.
- We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.
  - Data Columns obtained by using SpaceX RESP API:
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
  - Data Columns obtained by using Wikipedia Web Scraping:
    - Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

# Data Collection – SpaceX API




GitHub URL

# Data Collection – Web Scraping






GitHub URL

# Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.
- In this lab we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

Perform Exploratory Data Analysis and determine training labels

↓

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome of the orbits

Create a landing outcome label from Outcome column

Exporting the data to CSV



GitHub URL

# EDA with Data Visualization

- Charts plotted:
    - Flight Number vs. Payload Mass, Flight Number vs Launch Site, Payload Mass vs Launch Site, Success Rate by Orbit Type,  Flight Number vs Orbit Type, Payload Mass vs Orbit type, Launch Success Yearly Trend
- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series).



GitHub URL

# EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- 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
- 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



[GitHub URL]

# 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 its latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Colored Markers of the launch outcomes for each Launch Site:
  - Added colored Markers of successful (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site and its proximities:
  - Added colored Lines to show distances between the Launch Site KSC LC-39A and its proximities like Railway, Highway, Coastline and Closest City.



[GitHub URL]

# Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Successful Launches (All Sites/Specific Site):
  - Added a pie chart to show the total successful launch counts for all sites and the Success vs Failed counts for the site, if a specific 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 different Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success.



GitHub URL

# Predictive Analysis (Classification)

Create a NumPy array from the column "Class" in data

Standardize the data with StandardScaler(), then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and kNN models

Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

GitHub URL

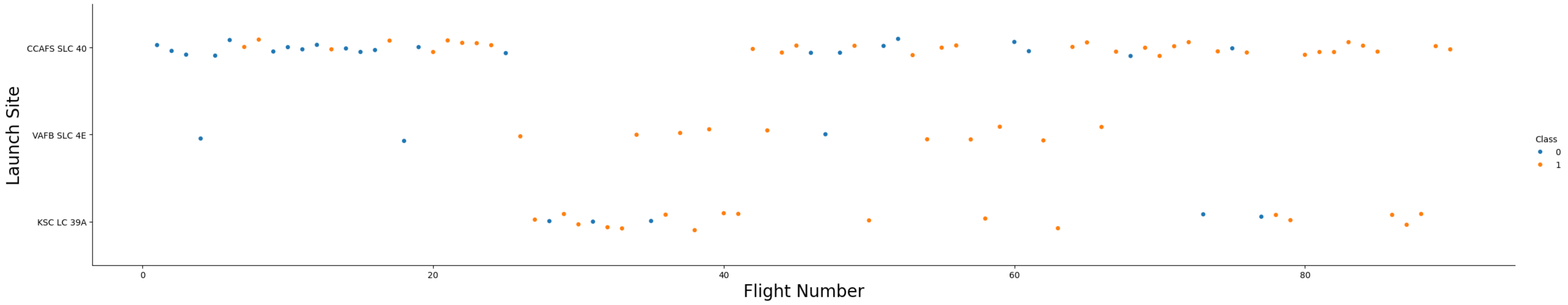
# Results

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

Section 2

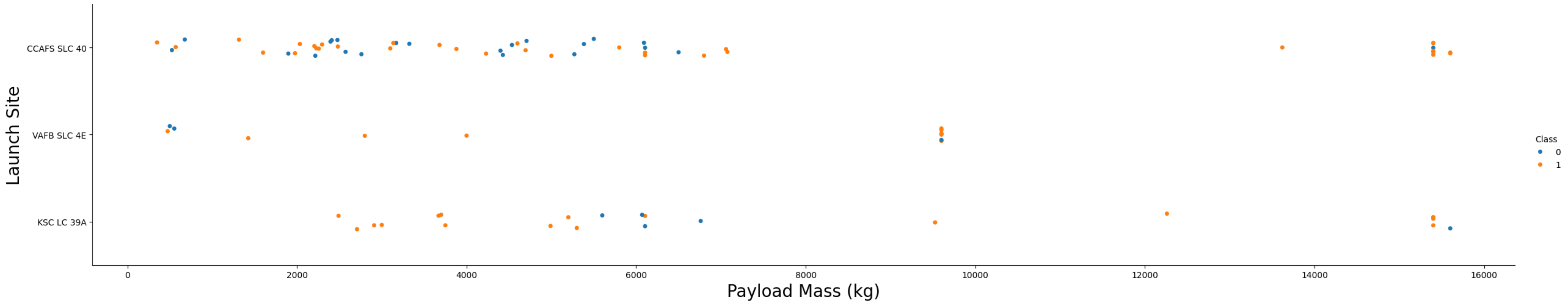
# Insights drawn from EDA

# Flight Number vs. Launch Site



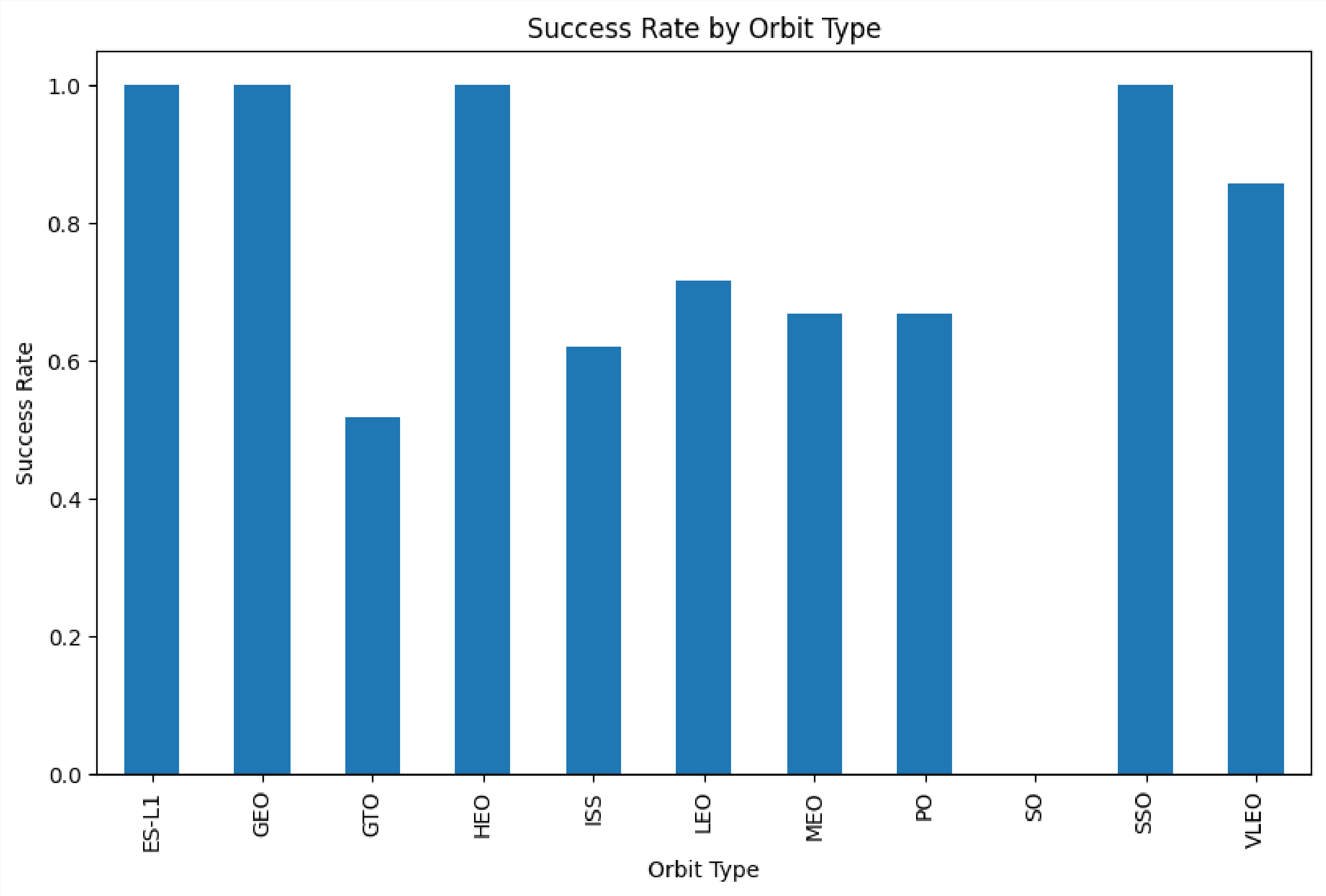

- Successes are more prevalent for the most recent flights as compared to the earliest flights, across all launch sites.
    - While the 5 earliest flights at launch site CCAFS SLC-40 were failures, the most recent 9 were successes
    - The 5 most recent flights at VAFB SLC-4E and KSC LC-39A were all successes
- One of the most valuable aspects of Exploratory Data Analysis (EDA) is the ability to prompt questions for further investigation. For example:
    - Why did flights at CCAFS SLC-40 cease around flight number 30 and resume just after flight number 40?
    - Is this related to flights at KSC LC-39A resuming and ending over the same range?

# Payload vs. Launch Site



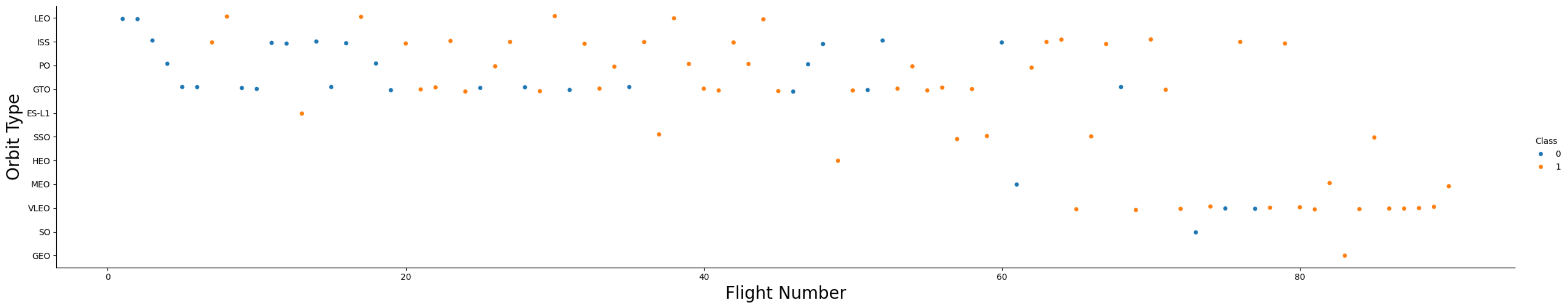

- Similar to the prior slide, this visual raises more questions for me than explanations. Again, this is an important element of EDA so I think this is a positive characteristic:
  - Is there a reason why flights with a Payload Mass less than 3,000 kg didn't occur at launch site KSC LC 39A?
  - Similarly, is there an explanation for why flights with a Payload Mass greater than 10,000 kg didn't occur at launch site VAFB SLC 4E?
  - Why is there a noticeable gap in flights with a Payload Mass between ~7,000 kg and ~14,000 kg at launch site CCAFS SLC 40?

# Success Rate vs. Orbit Type



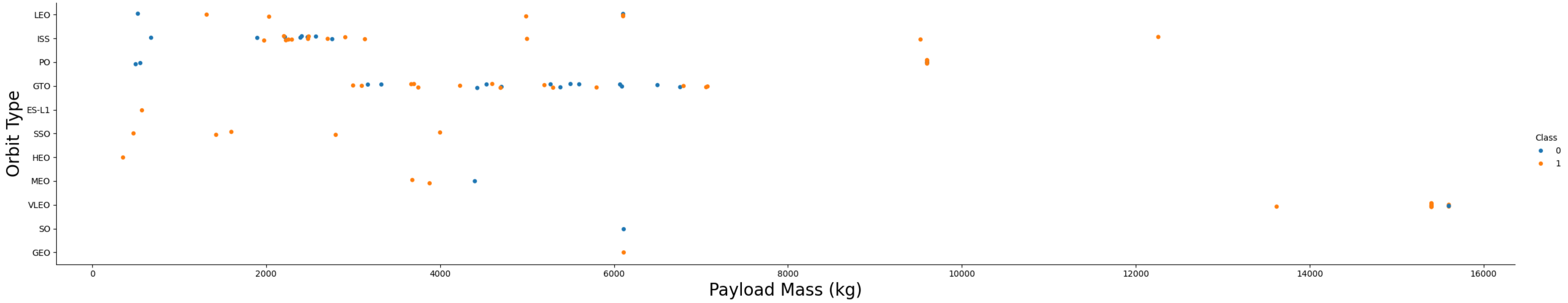

- 100% success rate for flights launched into ES-L1, GEO, HEO, and SSO
- However, it is important to note the sample sizes are extremely small; success rates for ES-L1, GEO, and HEO based on a single flight
- VLEO is noteworthy in that 12 of the 14 flights launched into VLEO were successful; 86% success rate

# Flight Number vs. Orbit Type



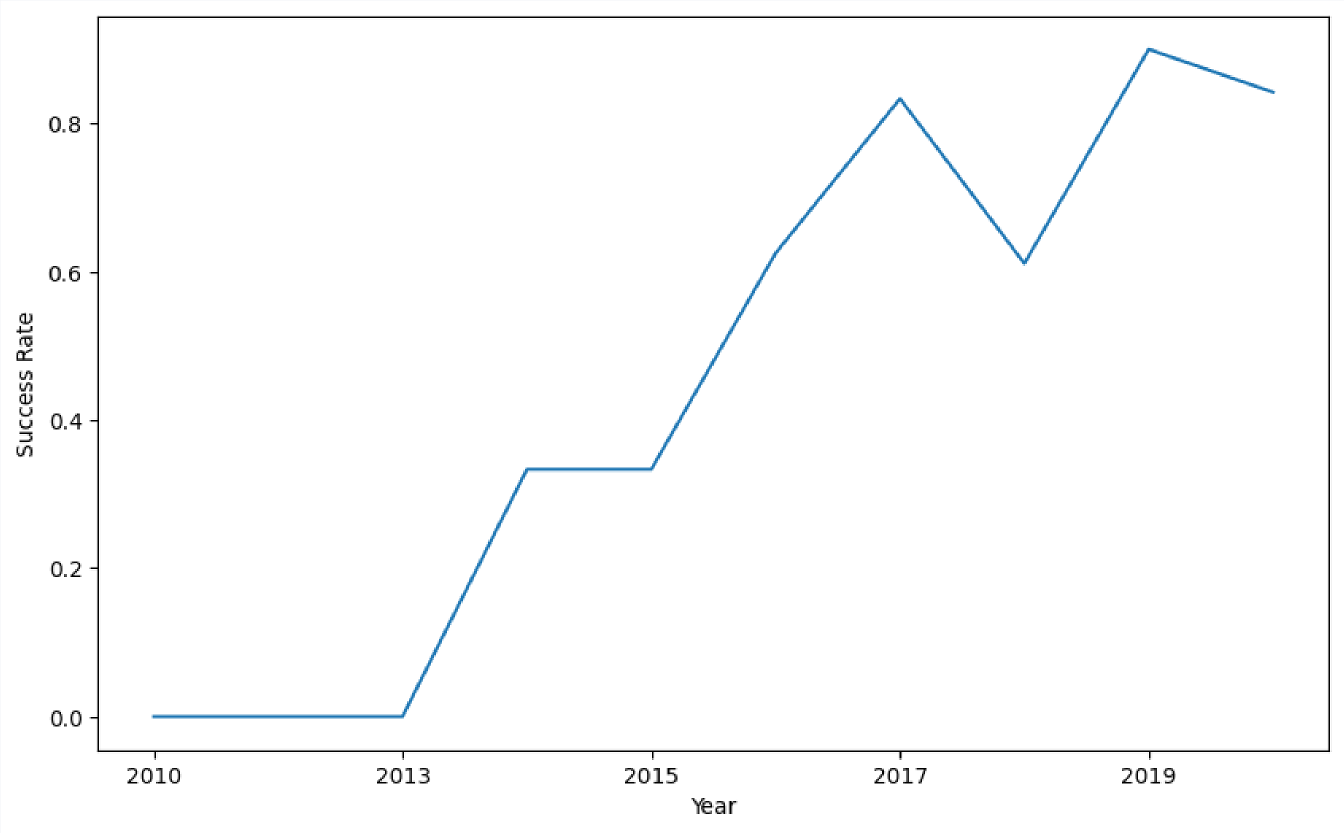

- In the LEO orbit, the Success appears related to the number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit
- Is there a reason why orbits starting at MEO (and to a lesser degree SSO) seem to be strictly associated with larger flight numbers?

# Payload vs. Orbit Type




- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.
- Noticeable contrast in success rates between orbits LEO, ISS, PO (failures) and ES-L1, SSO, HEO (successes) for the lightest payloads (<1,000 kg)

# Launch Success Yearly Trend




- The success rate since 2013 kept increasing until 2017 (stable in 2014) and after 2015 it started increasing

# All Launch Site Names

```
%sql select distinct "Launch_Site" as "Launch Site" from SPACEXTABLE;
```

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

Explanation:

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

# Launch Site Names Begin with 'CCA'

%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5;

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 6/4/2010 | 0.8 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0.0 | LEO | SpaceX | Success | Failure (parachute) |
| 12/8/2010 | 0.7 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0.0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 5/22/2012 | 0.3 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525.0 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 10/8/2012 | 0.0 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500.0 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 3/1/2013 | 0.6 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677.0 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Total Payload Mass

```
%sql select sum("PAYLOAD_MASS__KG_") as "Total Payload Mass (kg)" from SPACEXTABLE where "Customer" = 'NASA (CRS)';
```

| Total Payload Mass (kg) |
| --- |
| 45,596 |

Explanation:

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

# Average Payload Mass by F9 v1.1

```
%sql select avg("PAYLOAD_MASS_KG_") as "Average Payload Mass (kg)" from SPACEXTABLE where "Booster_Version" = 'F9 v1.1';
```

| Average Payload Mass (kg) |
|---|
| 0 |

Explanation:

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

# First Successful Ground Landing Date

```
%sql select min("Date") as "Date" from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)";
```

| Date |
|---|
| 12/22/2015 |

Explanation:

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

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select distinct "Booster_Version" as "Booster Version" from SPACEXTABLE where "Landing_Outcome" = "Success (drone ship)"
and "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000;
```

| Booster Version |
|---|
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

Explanation:

- List 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

```
%sql select Mission_Outcome as "Mission Outcome", count(*) as "Count" from SPACEXTABLE group by Mission_Outcome;
```

| Mission Outcome | Count |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

Explanation:

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

# Boosters Carried Maximum Payload

```
%sql select distinct "Booster_Version" as "Booster Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE);
```

| 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 |

Explanation:

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

# 2015 Launch Records

```
%sql select substr("Date", 6, 2) as Month, "Landing_Outcome" as "Landing Outcome", "Booster_Version" as "Booster Version", "Launch_Site" as "Launch Site" from SPACEXTABLE where "Landing_Outcome" = "Failure (drone ship)" and substr("Date", 0, 5) = '2015';
```

| Month | Landing Outcome | Booster Version | Launch Site |
|---|---|---|---|
| 1 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 4 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

Explanation:

- 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.

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

```
%sql select "Landing_Outcome" as "Landing Outcome", count(*) as "Count" from SPACEXTABLE where "Date" between '2010-06-04' and '2017-03-20' group by "Landing_Outcome" order by Count desc;
```

| 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 |

Explanation:

- 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.

Section 3

# Launch Sites Proximities Analysis

# All launch sites' location markers on a global map

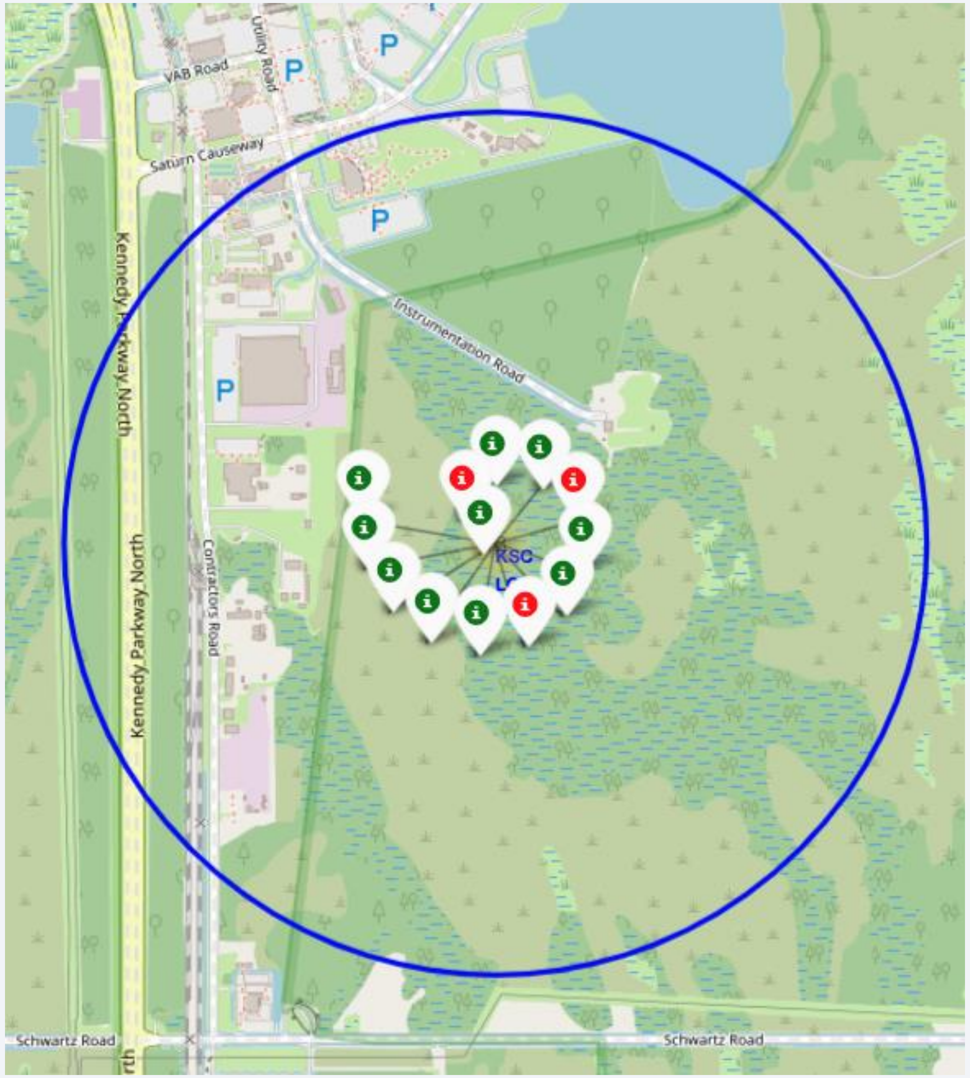
Explanation:

- CCAFS SLC-40/CCAFS LC-40
  - Chosen for its proximity to the Atlantic Ocean, allowing rockets to launch eastward and take advantage of Earth's rotation for added velocity. The ocean also provides a safe area for spent rocket stages to fall.
- KSC LC-39A
  - LC-39A was originally built for the Apollo program and later used for the Space Shuttle. Its location on the East Coast allows for efficient launches into equatorial orbits, which are ideal for missions to the International Space Station (ISS) and geostationary orbits.
- VAFB SLC-4E
  - Vandenberg's location is ideal for launching satellites into polar or sun-synchronous orbits, as rockets can launch southward over the Pacific Ocean without overflying populated areas. This is in contrast to the eastward launches from Cape Canaveral.

# Color-labeled launch records on the map

Explanation:

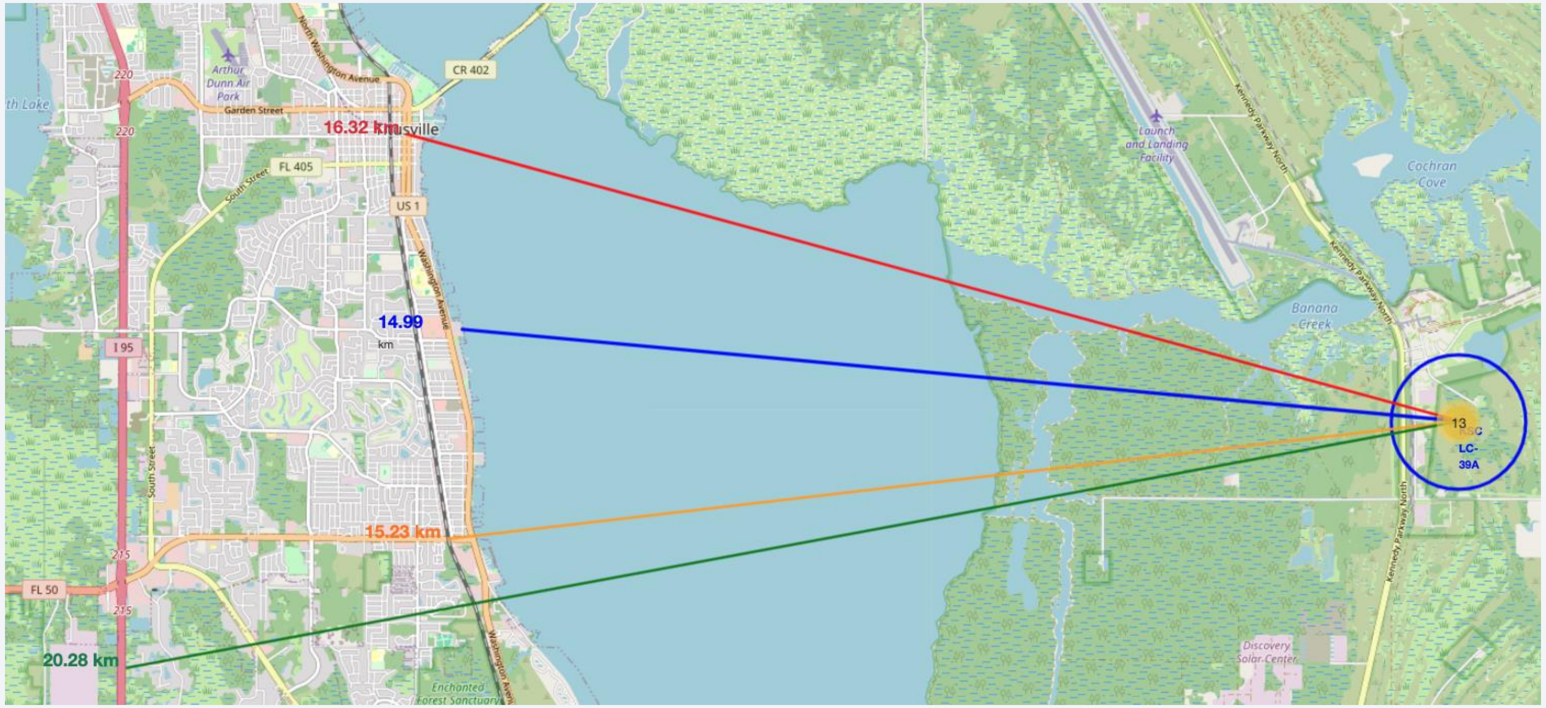
- From the color-labeled markers we can easily identify which launch sites have higher (and lower) success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A displayed in screenshot:
  - 76.9% success rate (highest of all launch sites)



# Distance from launch site KSC LC-39A to its proximities

Explanation:

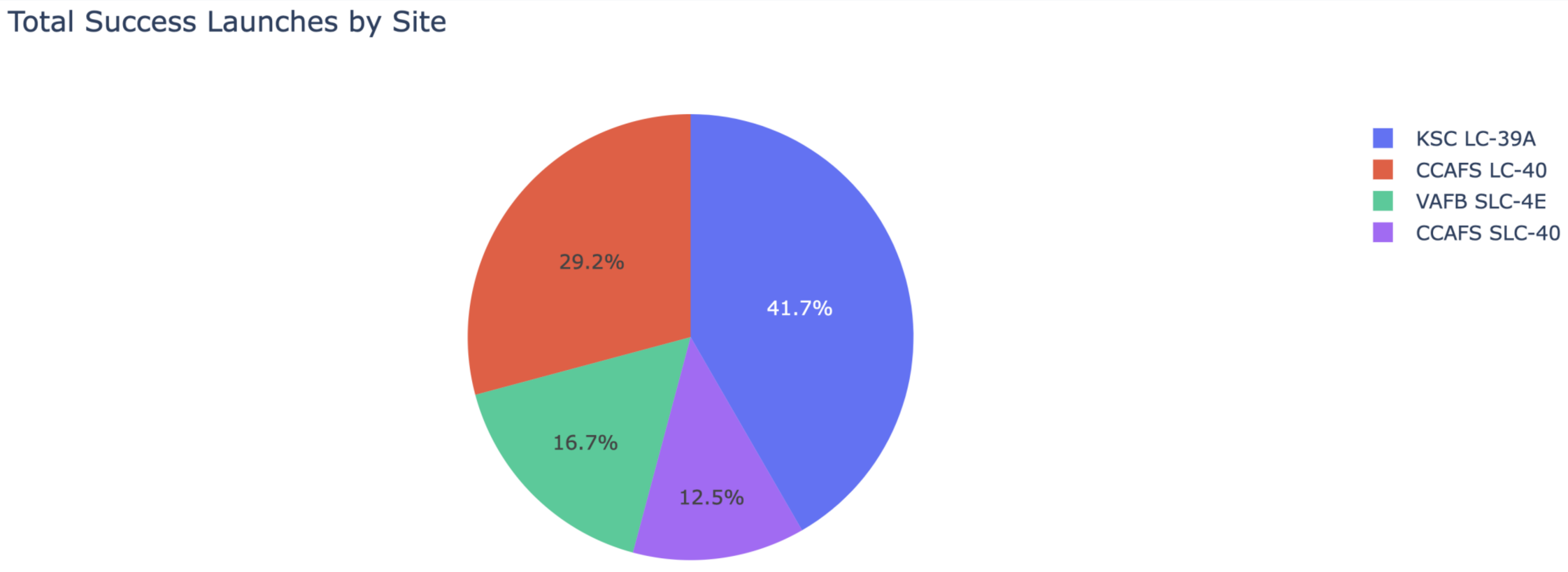
- From the visual analysis of launch site KSC LC-39A we can clearly see that it is:
  - 14.99 km (9.31 miles) to coastline
  - 15.23 km (9.46 miles) to railway
  - 16.32 km (10.14 miles) to city of Titusville
  - 20.28 km (12.60 miles) to highway



Section 4

# Build a Dashboard with Plotly Dash

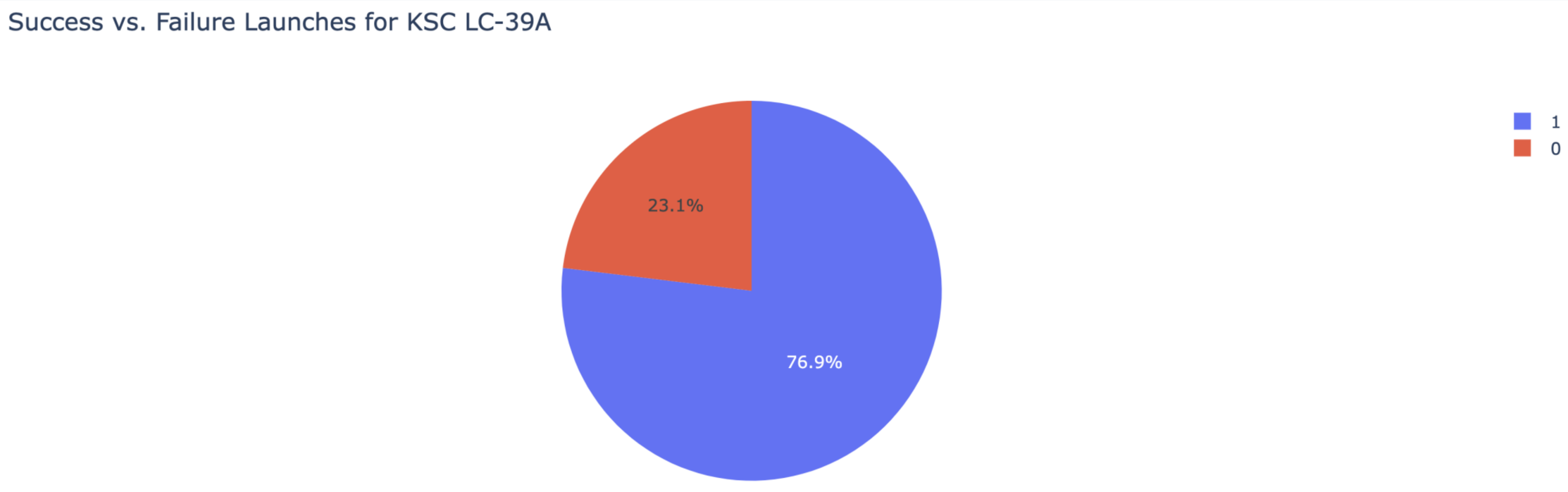
# Launch Success for All Sites




Explanation:

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

# Launch Site with Highest Launch Success Rate



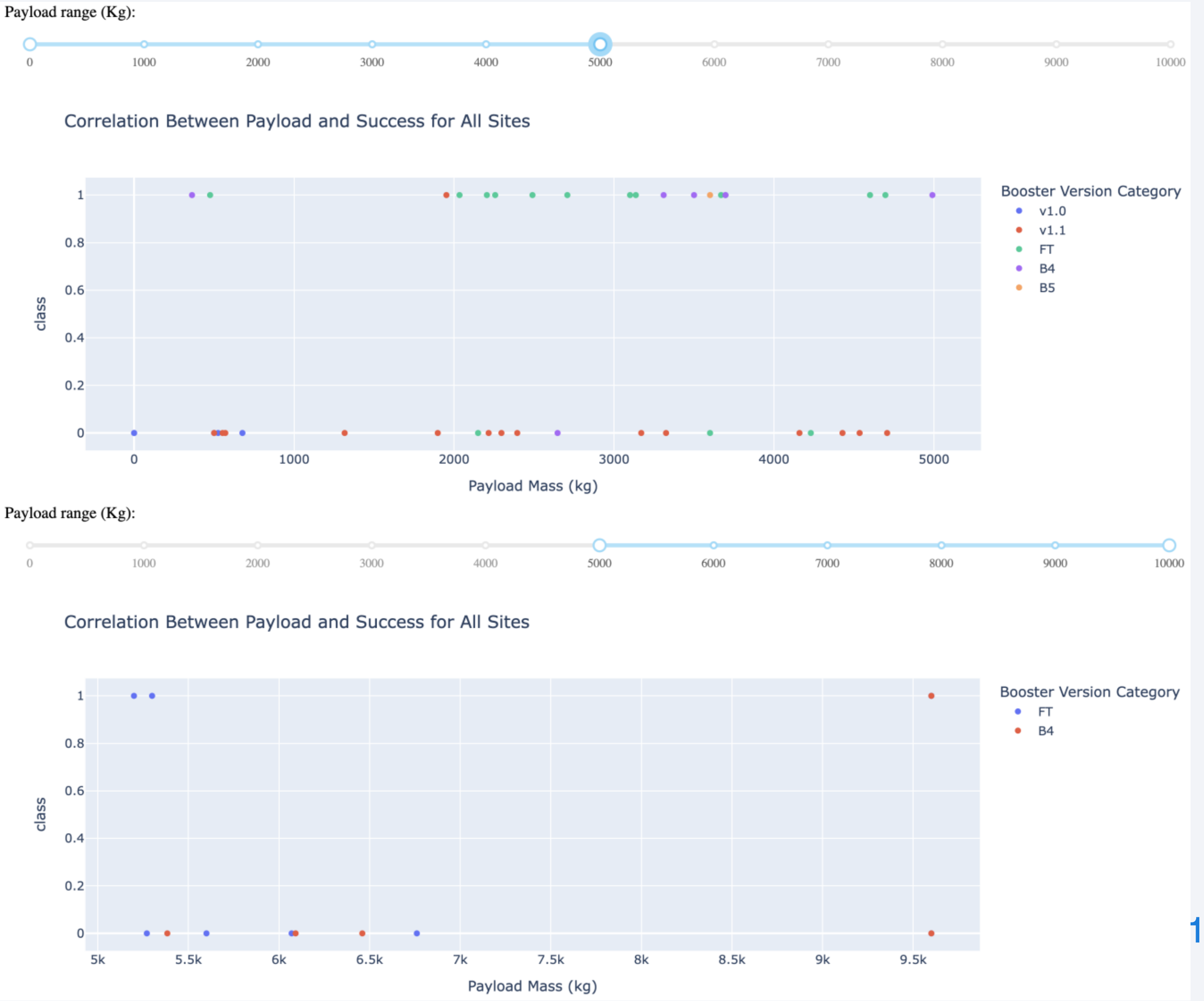

Explanation:

- 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

Explanation:

- The chart shows that payloads between 2,000 kg and 5,500 kg generally have higher success rates
- Success rates are highest in the 3,000 kg to 4,000 kg payload range
- Booster version FT flights are especially successful in the 2,000 kg to 3,000 kg payload range
- Booster version 1.1 flights have higher failure rates across all payload ranges



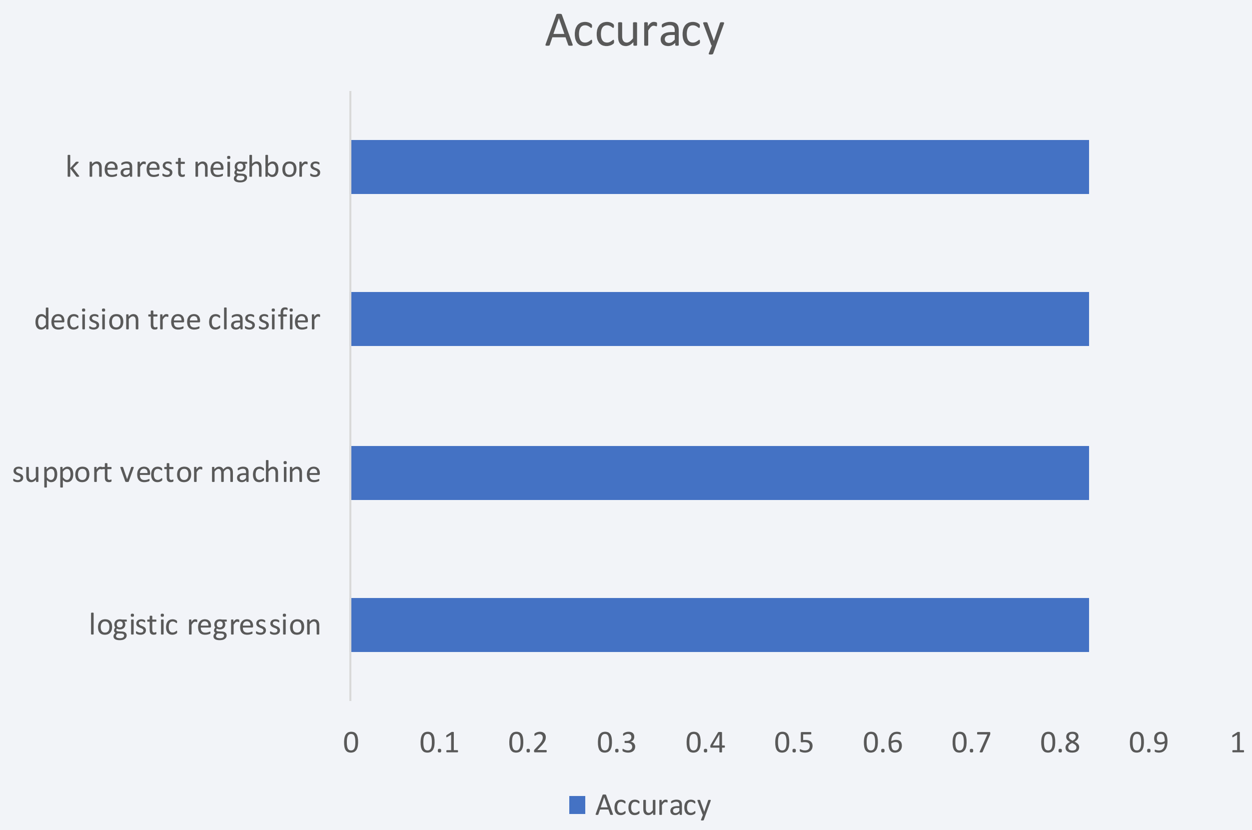
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

Explanation:

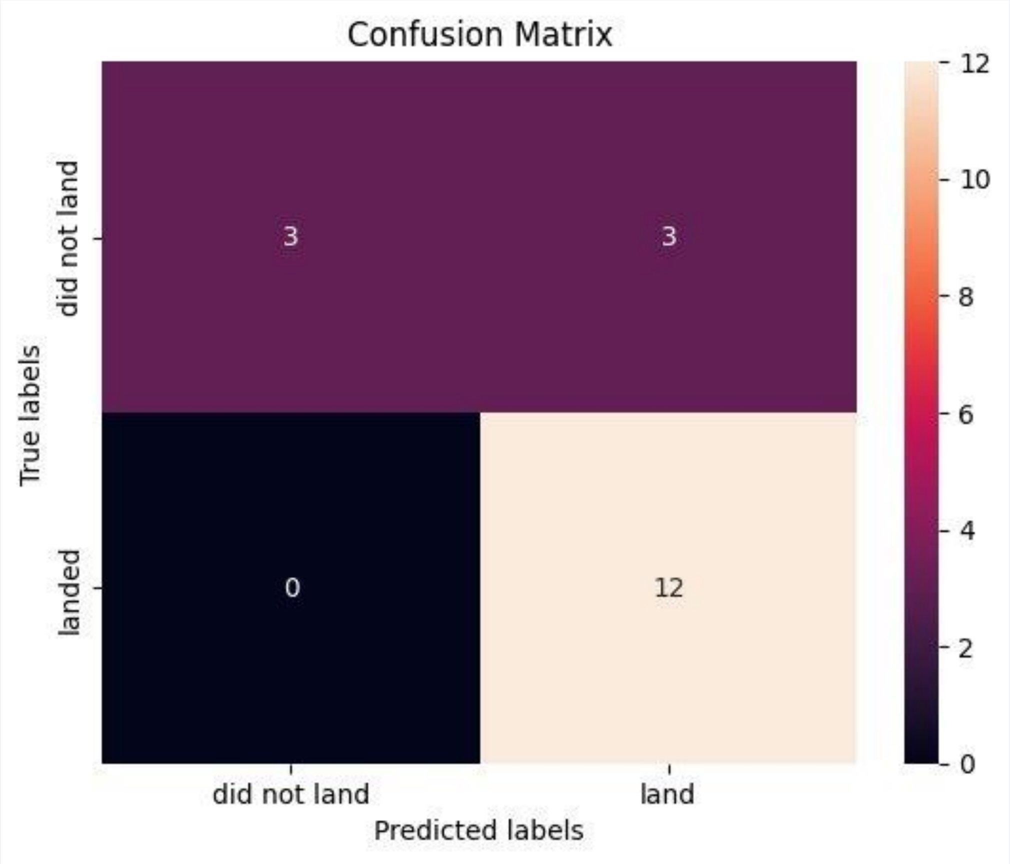
- The test set Accuracy across all models is 0.833; we cannot confirm which method performs best.



# Confusion Matrix

Explanation:

- Of the 15 flights the model predicts will land successfully, 12 landed successfully
    - This represents model precision (80%)
- Of the12 flights that landed successfully, the model predicts all 12 will land successfully
    - This represents model recall (100%)
- The model has room for improvement with respect to false positives (top-right quadrant)



# Conclusions

- Based on Accuracy metric alone, we cannot determine which model is best.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of the launch sites are in close proximity to the Equator line and all the sites are in very close proximity to the coastline.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest launch success rate across all the sites.

# Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!