
# Winning Space Race with Data Science

Mary-PN

May 20th, 2022

# Outline

# Executive Summary

Conducting thorough market research is an important part of crafting a business plan. Our competitor, SpaceX has dominated the market by simply reusing the first stage of the rockets after a successful landing.

We have collected publicly available data on SpaceX Falcon 9 rocket launches, gained some insights,  and built machine learning models to predict whether the first stage will land successfully. This was achieved though the following:

- Data Collection
- Data Analysis
- Predictive Modelling using various machine learning algorithms

The outcome provides business insights resulting in becoming a competitor to SpaceX!

# Introduction

The main objective of this data analysis is to provide insight to the company's decision makers to compete with SpaceX.

SpaceX 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.

We have trained a machine learning model and used public information to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully. This will mean the company will reuse the first stage and reduce cost.

Section 1

# Methodology

# Methodology

**Data collection methodology:**

- Making requests to the SpaceX API
- Web scraping to collect Falcon 9 historical launch records on Wikipedia

**Perform data wrangling**

- Data cleaning to create consistent and technically correct format and values

**Perform exploratory data analysis (EDA) using visualization and SQL**

**Perform interactive visual analytics using Folium and Plotly Dash**

**Perform predictive analysis using classification models**

- Build various supervised learning models, evaluate the models and determine the best model to predict the outcome of the first stage landing

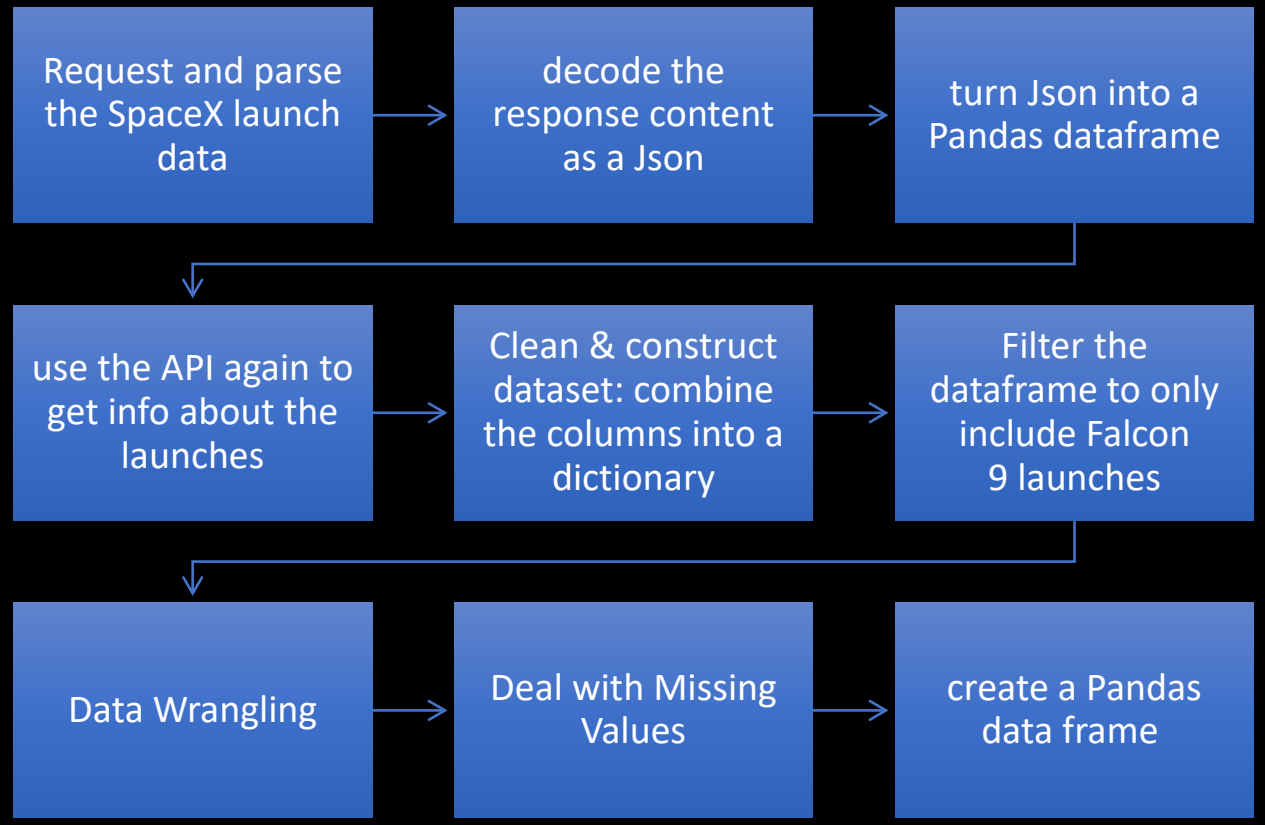
# Data Collection

The main purpose of data collection is to gather information in a measured and systematic manner to ensure accuracy and facilitate data analysis. Since the data collected is meant to provide content for data analysis, the information gathered must be of the highest quality for it to be of value. We used the following methods which are cost effective:

- SpaceX REST API request
- Web scraping related Wiki pages

We will look at the process used in each method in more details.

# Data Collection – SpaceX API

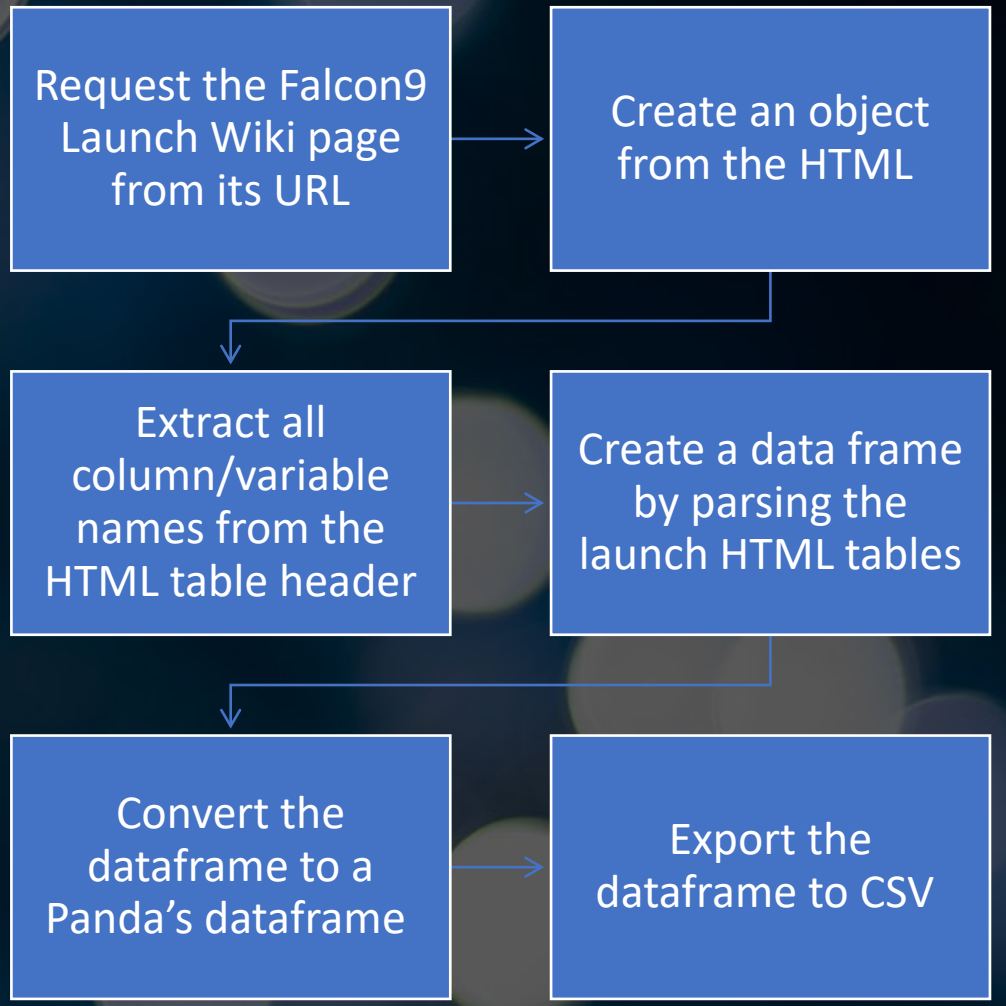



SpaceX REST API request provided data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

Reference URL: https://github.com/Mary-PN/IBM_Capstone_Project/blob/main/Data%20Collection%20API%20Lab.ipynb

# Data Collection: Scraping

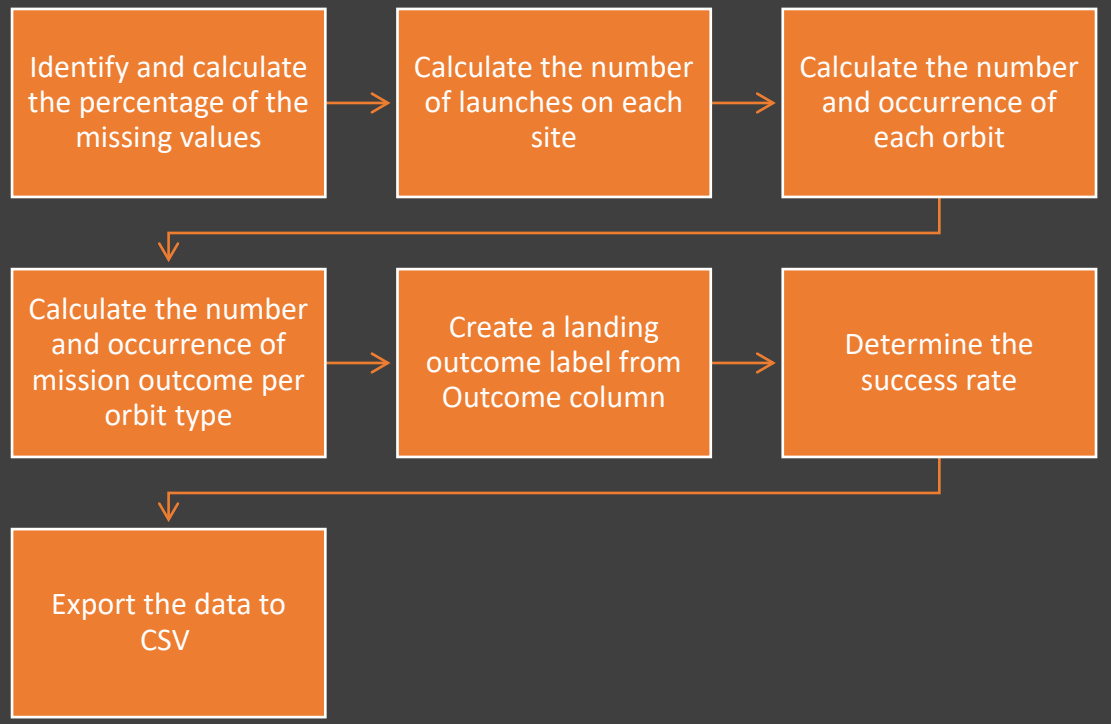
Web scraping related Wiki pages provided tables that contain valuable Falcon 9 and Falcon Heavy Launches Records.




Reference URL: https://github.com/Mary-PN/IBM_Capstone_Project/blob/main/Data%20Collection%20with%20Web%20Scraping%20Lab.ipynb

# Data Wrangling

In this section we conducted Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.



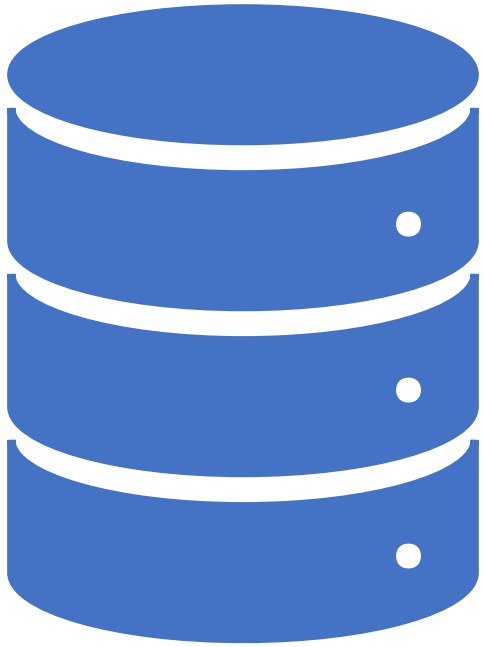

Reference URL: https://github.com/Mary-PN/IBM_Capstone_Project/blob/main/EDA.ipynb

# EDA with Data Visualization

EDA with Data Visualization is the process of describing the data by means of statistical and visualization techniques in order to bring important aspects of that data into focus for further analysis. We used the following:

- Visualized the relationship between Flight Number and Launch Site using scatterplot
- Visualized the relationship between Payload and Launch Site using scatterplot
- Visualized the relationship between success rate of each orbit type using bar chart
- Visualized the relationship between Flight Number and Orbit type using scatterplot
- Visualized the relationship between Payload and Orbit type using scatterplot
- Visualized the launch success yearly trend

Reference URL: https://github.com/Mary-PN/IBM_Capstone_Project/blob/main/EDA-data-viz.ipynb



# EDA with SQL

Resuming the Exploratory Data Analysis, we performed the following SQL queries:

- Found the names of the unique launch sites in the space mission
- Found the total payload mass carried by boosters launched by NASA (CRS)
- Found the average payload mass carried by booster version F9 v1.1
- Displayed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Found the total number of successful and failure mission outcomes
- Found the names of the booster versions which have carried the maximum payload mass

Reference URL: https://github.com/Mary-PN/IBM_Capstone_Project/blob/main/EDA_with_SQL.ipynb

# Build an Interactive Map with Folium

Resuming the Exploratory Data Analysis, we performed the following using Folium:

- Marked all launch sites on map using Folium Markers
- Marker cluster was used to mark the success(Green)/failure(Red) launches for each site on the map: this enhanced the map by adding the launch outcomes for each site, and see which sites have high success rates.
- Added a MousePosition on the map to get coordinate for a mouse over a point on the map. As such, while you are exploring the map, you can easily find the coordinates of any points of interests (such as railway) to calculate the distances between a launch site to its proximities
- Polylines were added to connect the launch sites to any point of interest in the area

Reference URL: https://github.com/Mary-PN/IBM_Capstone_Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb


# Build a Dashboard with Plotly Dash

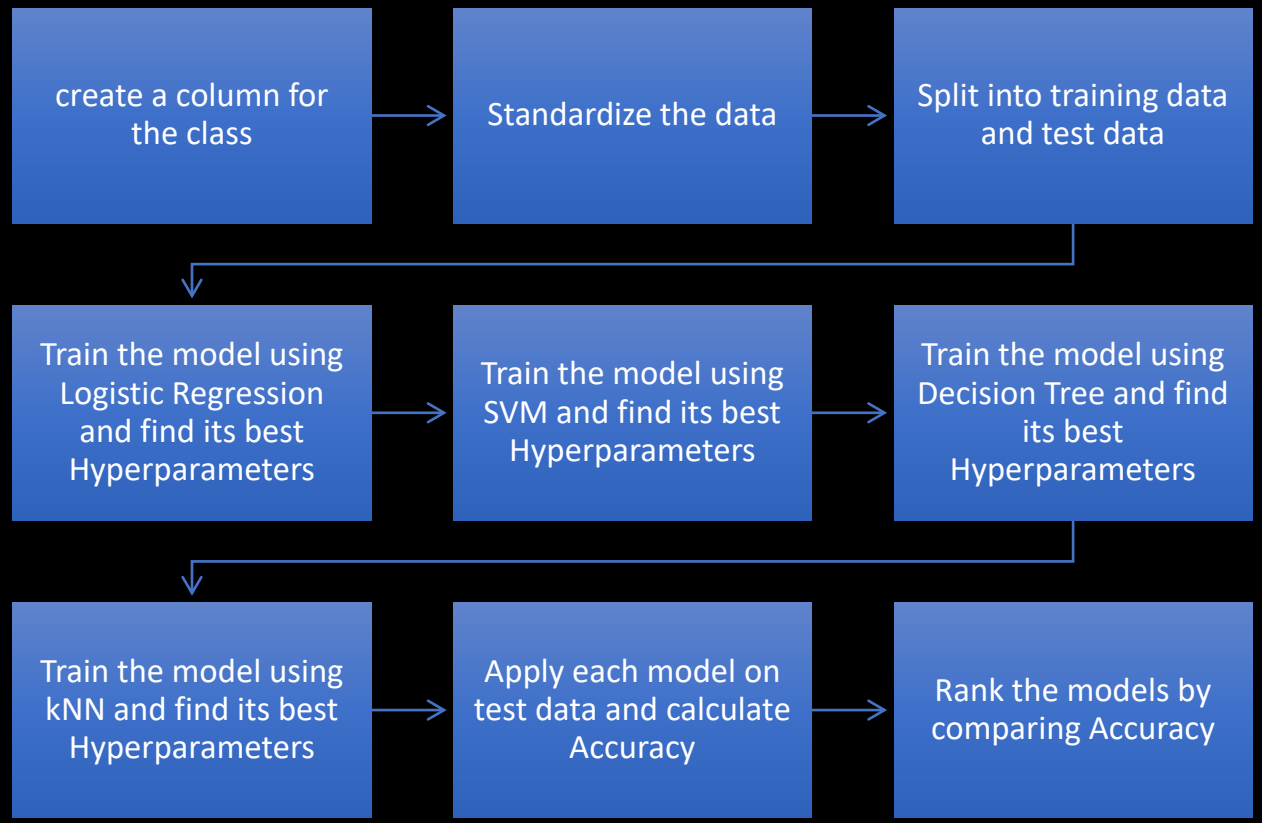
Using Plotly Dash an interactive dashboard was created demonstrating the following:

- A Pie Chart showing the success/failure rate at each launch site
- A Scatterplot demonstrating the success count on Payload Mass and Booster Version for all sites
- The charts can be drilled down further by selecting a specific site providing insights on the relationship between the Payload mass and Booster version with the success/ failure outcomes. Success is demonstrated by “1” and failure by “0” on each plot.

Reference URL: https://github.com/Mary-PN/IBM_Capstone_Project/blob/main/Final%20dashboard.py

# Predictive Analysis (Classification)

- Trained four Classification machine learning models using training set: Logistic Regression, SVM, Decision Tree, kNN.
- Best Hyperparameters were found using "*GridSearchCV*".
- Each model's performance was evaluated by the test data and generating confusion matrix
- The accuracy was calculated for each model
- Models were ranked based on the accuracy



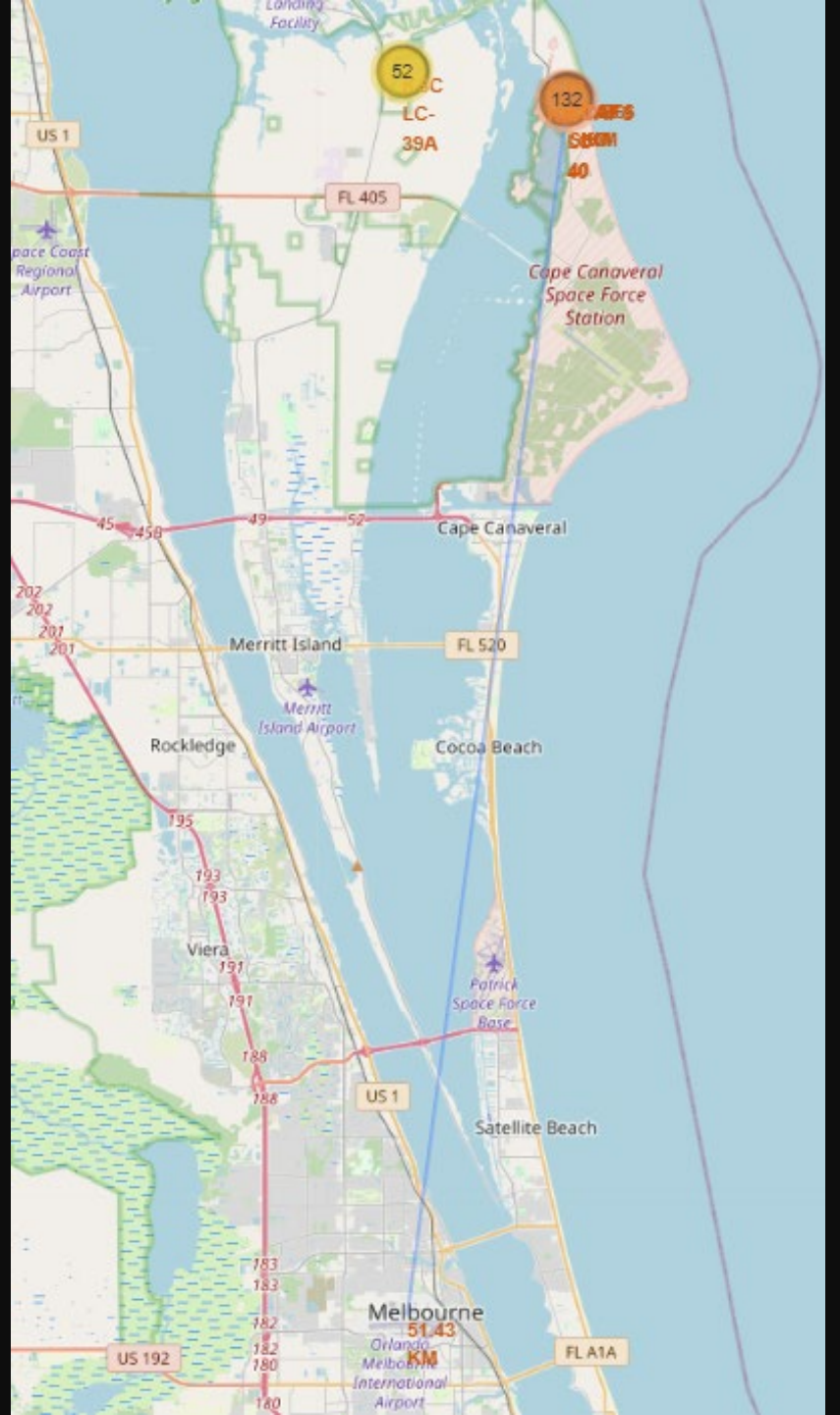

Reference URL: https://github.com/Mary-PN/IBM_Capstone_Project/blob/main/SpaceX_Machine%20Learning%20Predictive%20Analysis.ipynb

# Results

- Exploratory data analysis demonstrated that the success rate since 2013 kept increasing till 2020.
- 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.
- 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.
- Predictive analysis was conducted by building four machine learning models. kNN, SVM and Logistic regression models demonstrated an equal 83.3% accuracy achieved from the existing data. The decision Tree had a lower accuracy of 77.7%.

# Results

- The visual analytics revealed that:
- All launch sites are near railways, highways and coastlines. This consideration facilitates fast and flexible access to transportation.
- The sites are in a convenient proximity to airports.
- There is also safety consideration demonstrated through keeping the sites at a minimum distance of 50km from cities and crowded areas.

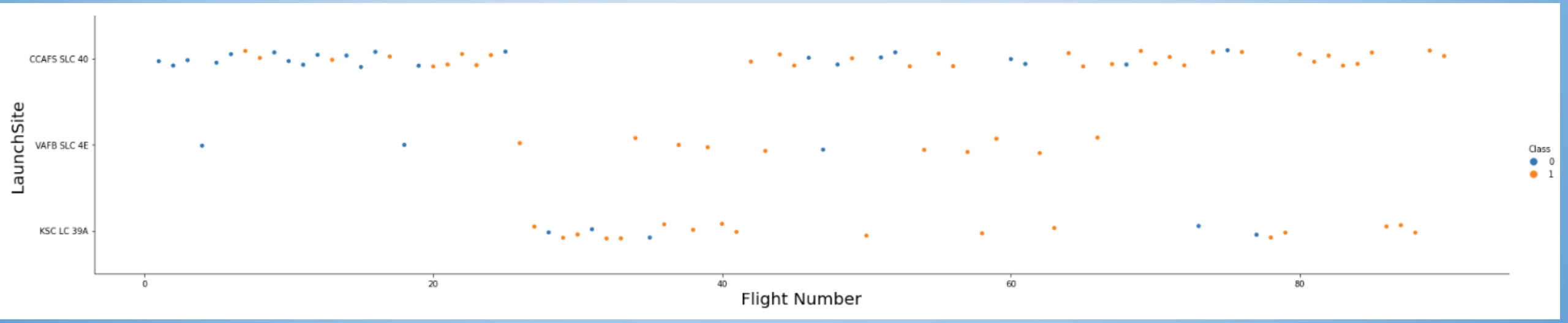


Section 2

# Insights drawn from EDA

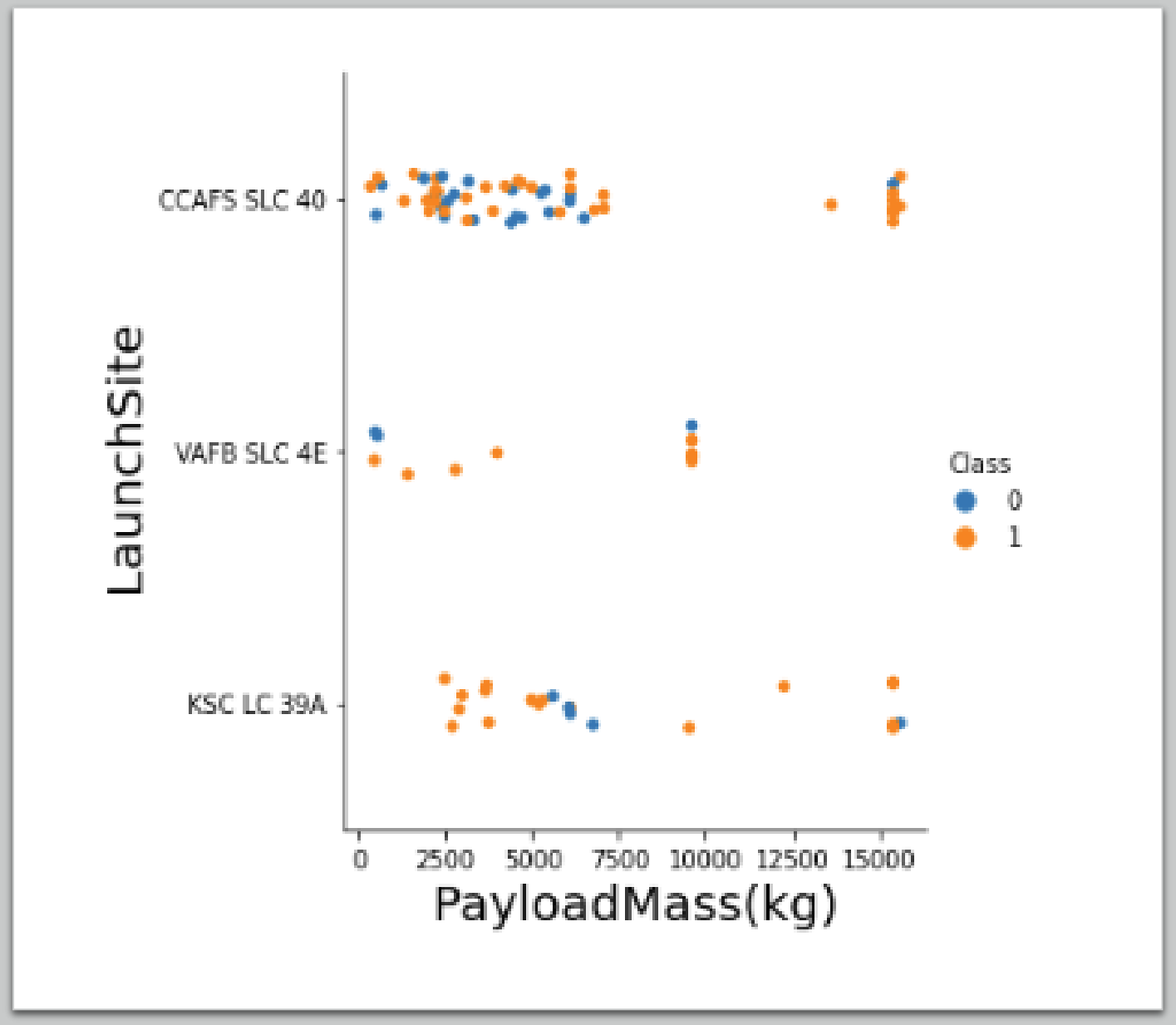
# Flight Number vs. Launch Site

- As the flight number increases the number of failures decreases. After approximately flight number=80, there are no failures.
- The CCAFS SLC 40 has highest number of lunches.
- There is no more than approximately flight number=70 at KSC LC 39A site.



# Payload vs. Launch Site

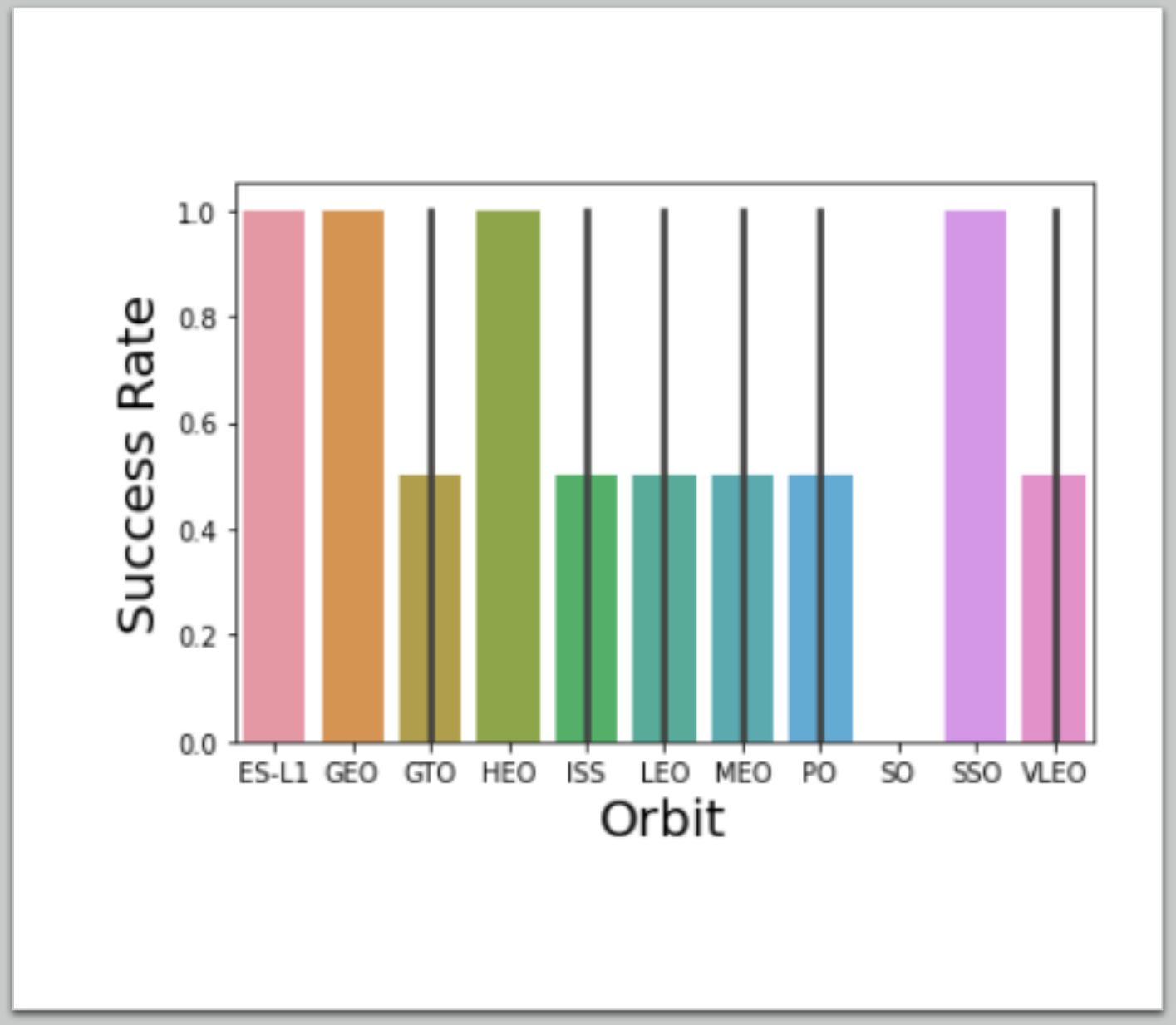
- As the Payload Mass increases the number of failures decreases.
- VAFB SLC 4E has no Payload Mass over 10000 Kg.



# Success Rate vs. Orbit Type

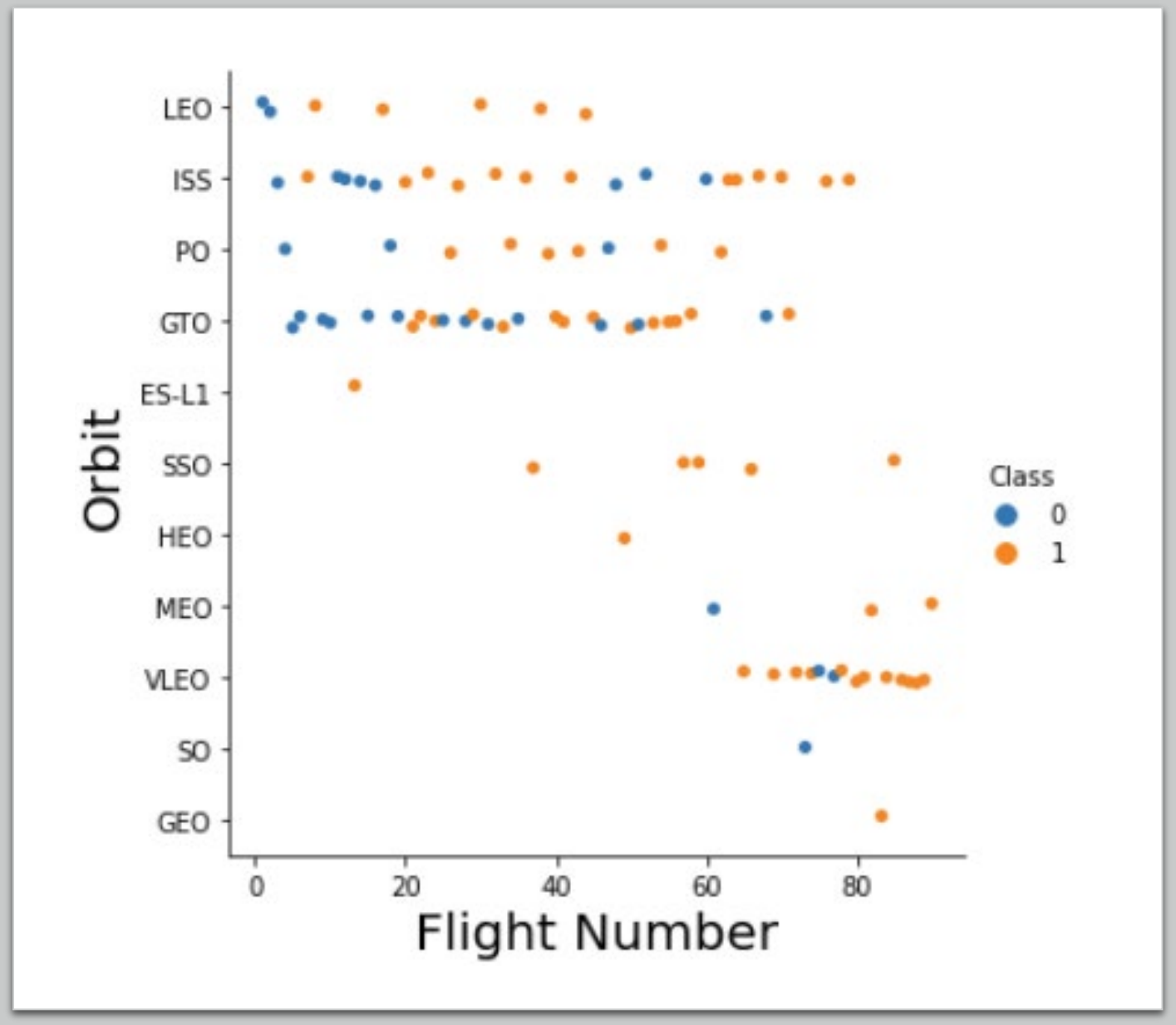
- The following four orbits have success rate of 100%:
  - ES-l1
  - GEO
  - HEO
  - SSO
- The SO has zero success rate, and the rest of the orbits have 50% success rate.

But we need to pay a closer attention to this at the next slide.



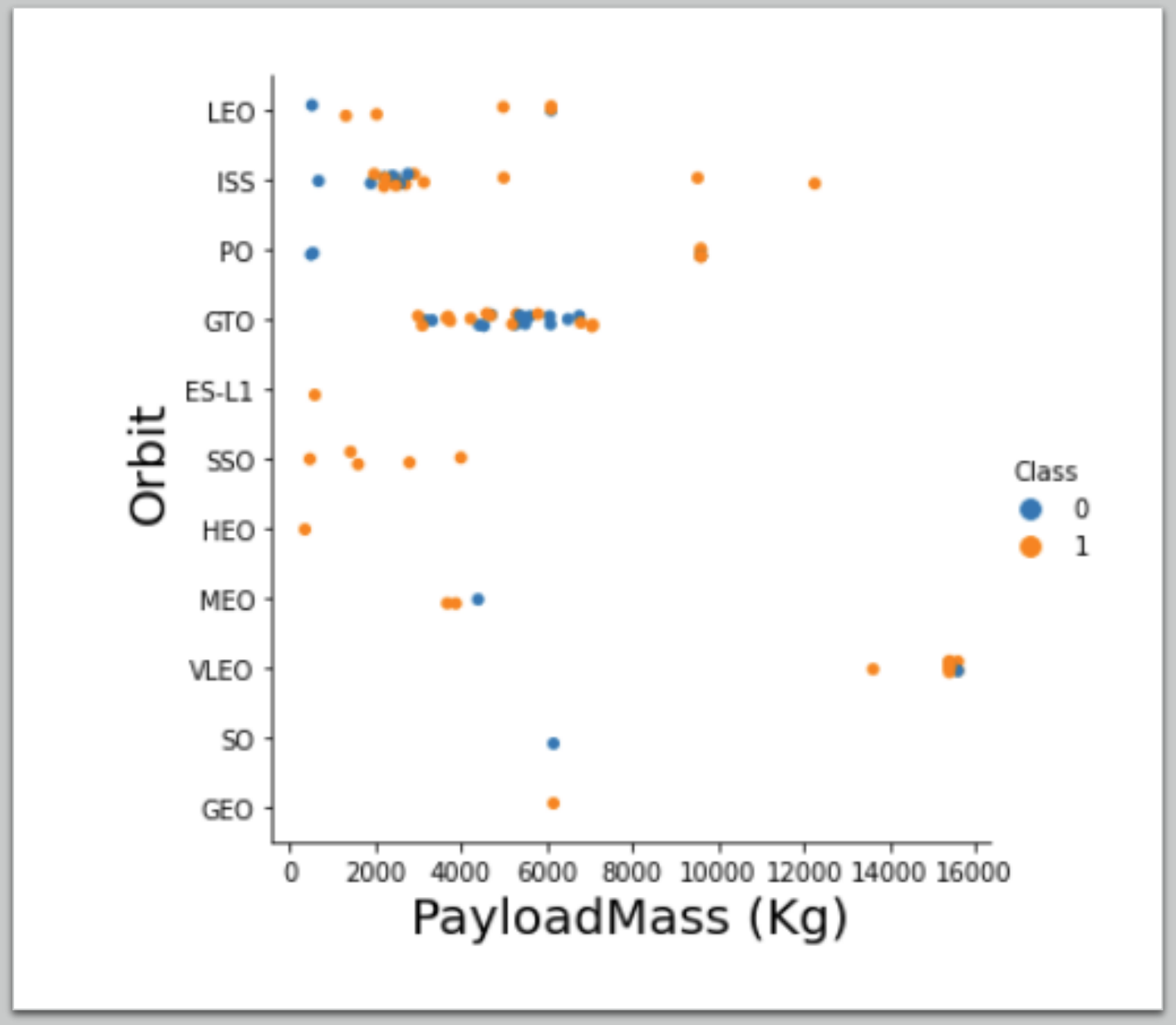
# Flight Number vs. Orbit Type

- It's important not to draw conclusions solely based on the previous slide as you notice the four orbits with highest success rate have also the lowest number of launches.
- In the LEO orbit the Success appears to be related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- Also as previously mentioned, as the flight number increases the number of failures decrease.



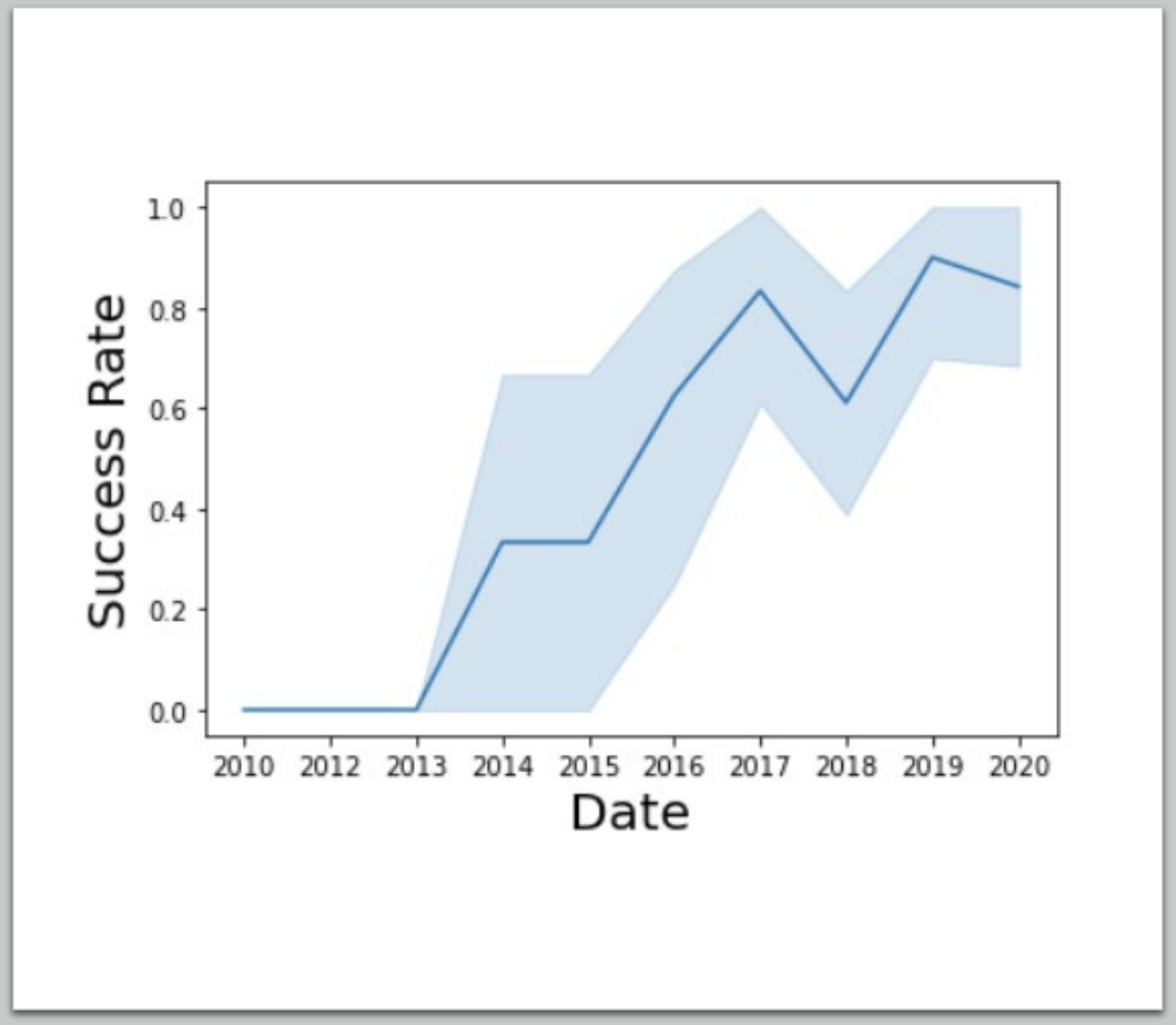
# Payload vs. Orbit Type

- Only 3 orbit types (ISS, PO, VLEO) have Payload Mass higher than 8000Kg which provides them with higher success rate.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are present.



# Launch Success Yearly Trend

- The success rate since 2013 has kept increasing till 2020!



# All Launch Site Names

The following are the names of the unique launch sites in the space mission in the gathered data:

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

# Launch Site Names Begin with 'CCA'

The followings are examples of 5 records where launch sites begin with `CCA`:

| 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

The total payload mass carried by boosters launched by NASA (CRS) are:

| customer | total_payload_mass |
|---|---|
| NASA (CRS) | 91192 |
| NASA (CRS), Kacific 1 | 5234 |

# Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is **2928 Kg**!

| average_payload_mass |
| --- |
| 2928 |

# First Successful Ground Landing Date

The date when the first successful landing outcome in ground pad was achieved was December 22nd, 2015!

| landing__outcome | first_success_date |
|---|---|
| Success (ground pad) | 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which had successfully landed on drone ship and had payload mass greater than 4000Kg but less than 6000Kg were:

| booster_version | payload_mass__kg_ | landing__outcome |
|---|---|---|
| F9 FT B1022 | 4696 | Success (drone ship) |
| F9 FT B1026 | 4600 | Success (drone ship) |
| F9 FT B1021.2 | 5300 | Success (drone ship) |
| F9 FT B1031.2 | 5200 | Success (drone ship) |
| F9 FT B1022 | 4696 | Success (drone ship) |
| F9 FT B1026 | 4600 | Success (drone ship) |
| F9 FT B1021.2 | 5300 | Success (drone ship) |
| F9 FT B1031.2 | 5200 | Success (drone ship) |

# Total Number of Successful and Failure Mission Outcomes

Below are the total number of successful and failure **mission outcomes**. Fortunately, the success rate is very high and impressive!

| mission_outcome | total |
|---|---|
| Failure (in flight) | 2 |
| Success | 198 |
| Success (payload status unclear) | 2 |

# Boosters Carried Maximum Payload

The following versions of boosters have carried the Maximum payload of 15600Kg

This information will be useful when high payload masses are being planned

| booster_version | payload_mass__kg_ |
|---|---|
| F9 B5 B1048.4 | 15600 |
| F9 B5 B1048.5 | 15600 |
| F9 B5 B1049.4 | 15600 |
| F9 B5 B1049.5 | 15600 |
| F9 B5 B1049.7 | 15600 |
| F9 B5 B1051.3 | 15600 |
| F9 B5 B1051.4 | 15600 |
| F9 B5 B1051.6 | 15600 |
| F9 B5 B1056.4 | 15600 |
| F9 B5 B1058.3 | 15600 |
| F9 B5 B1060.2 | 15600 |
| F9 B5 B1060.3 | 15600 |

# 2015 Launch Records

- This table provides information about the failed landing outcomes in drone ship, along with their booster versions, and launch site names for 2015

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

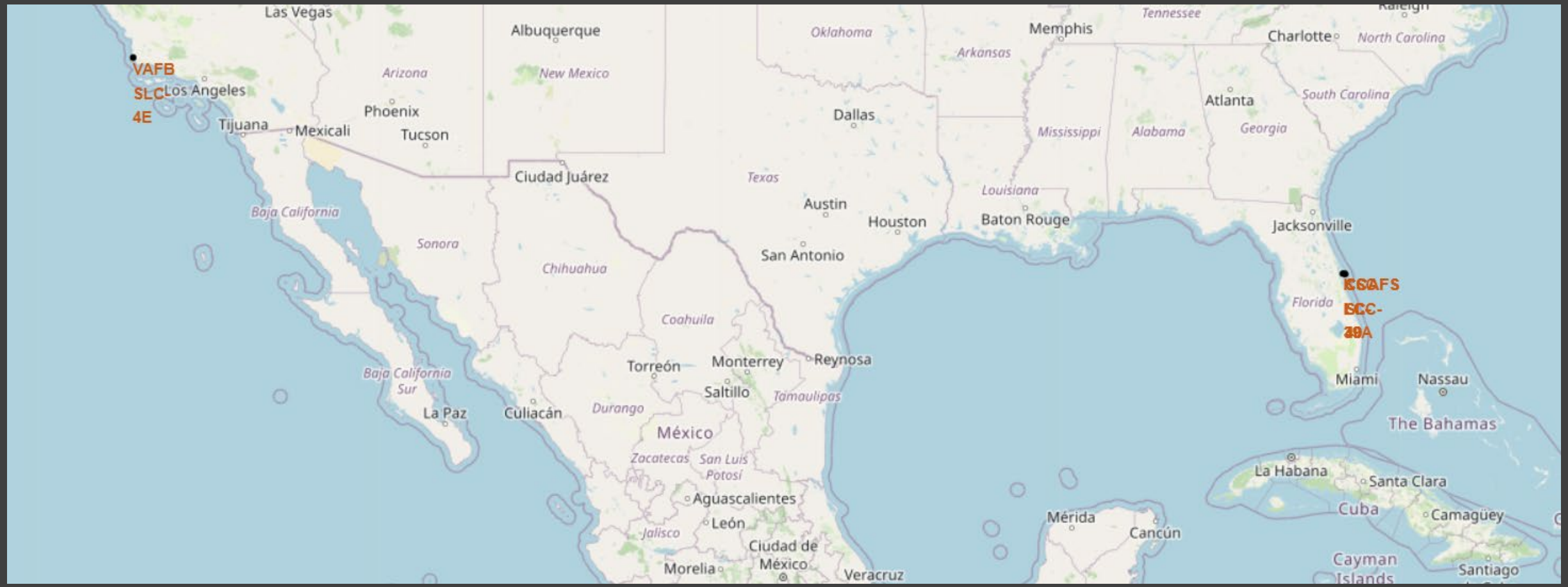
# Landing Outcomes Between 2010-06-04 and 2017-03-20

The following list is a ranked count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

| landing_outcome | total |
|---|---|
| No attempt | 20 |
| Failure (drone ship) | 10 |
| Success (drone ship) | 10 |
| Controlled (ocean) | 6 |
| Success (ground pad) | 6 |
| Failure (parachute) | 4 |
| Uncontrolled (ocean) | 4 |
| Precluded (drone ship) | 2 |

Section 3

# Launch Sites Proximities Analysis

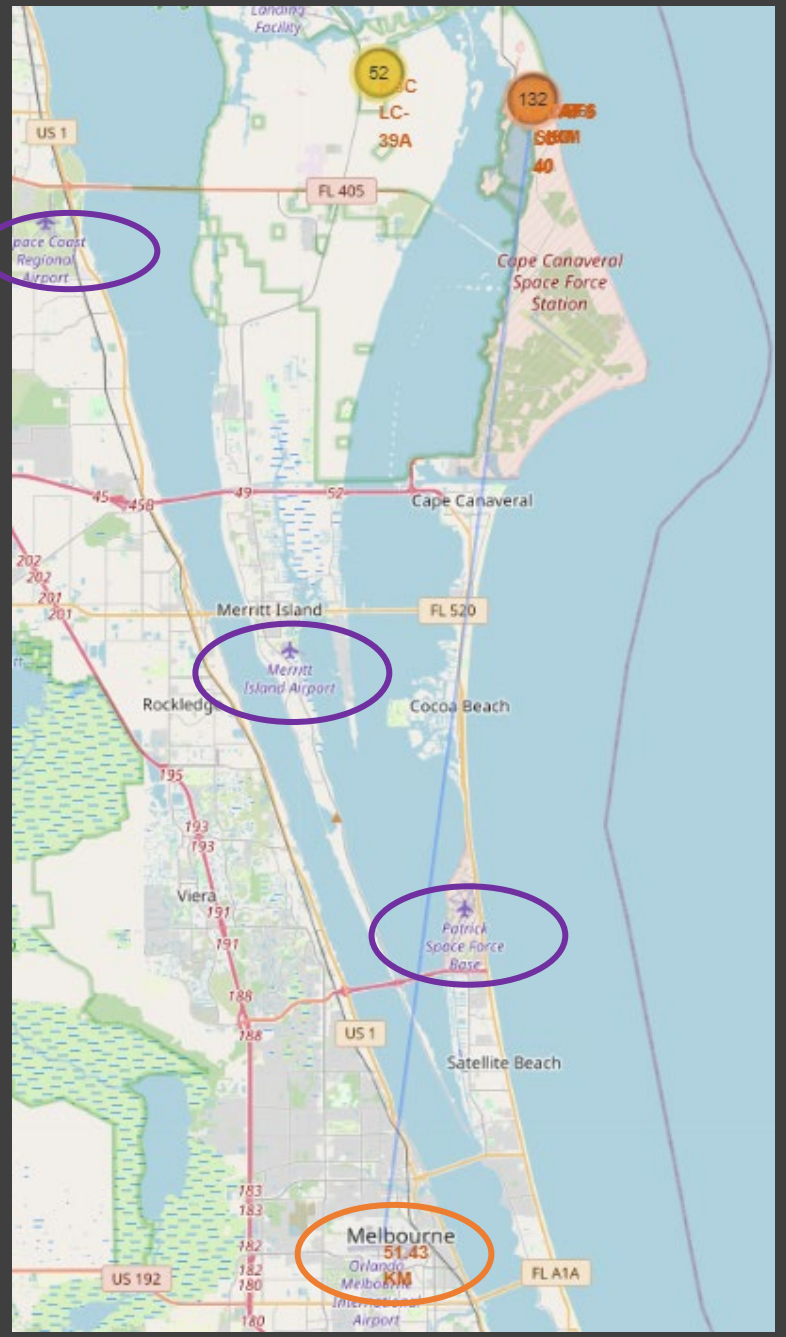


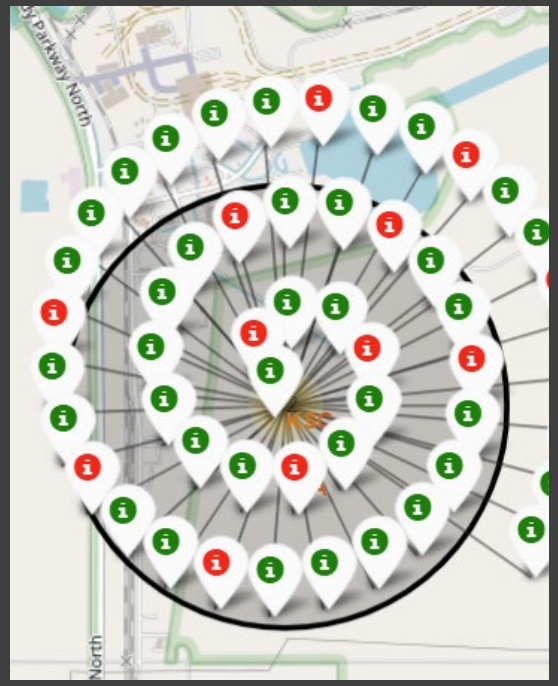

# Launch Sites on the Map

As we can see all lunch sites are near coasts, and in the outer parts of the country. Florida seems to be the most popular choice for launch sites, possibly due to central location while having a less severe weather conditions such as snow throughout the year.

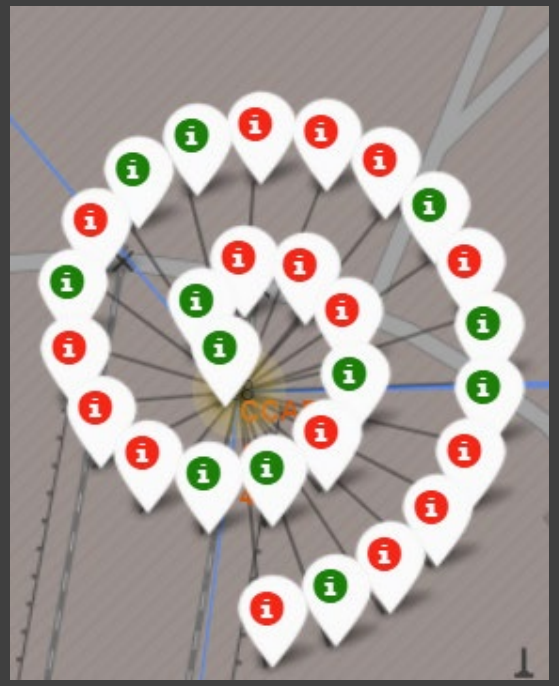
# Launch Site Surroundings

Beside being located on the coastline, the launch sites seems to be near major highways (yellow), Railways (black) , major roads (red), and airports (purple) . They also keep a reasonable distance from major cities (Melbourne 51.43 km distance)

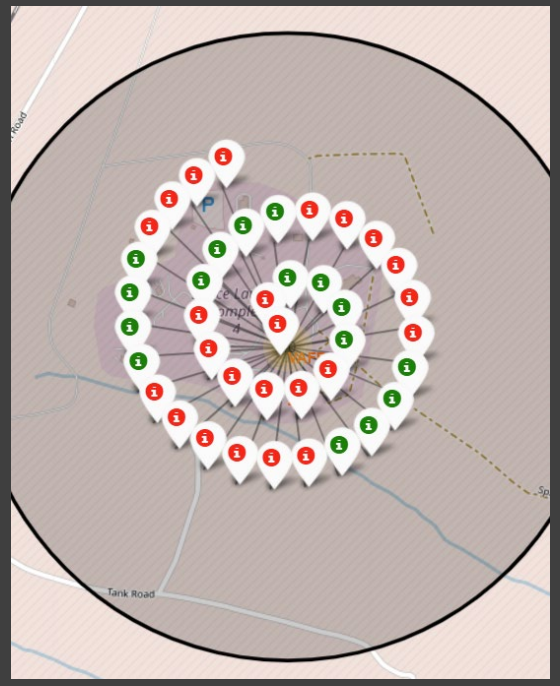




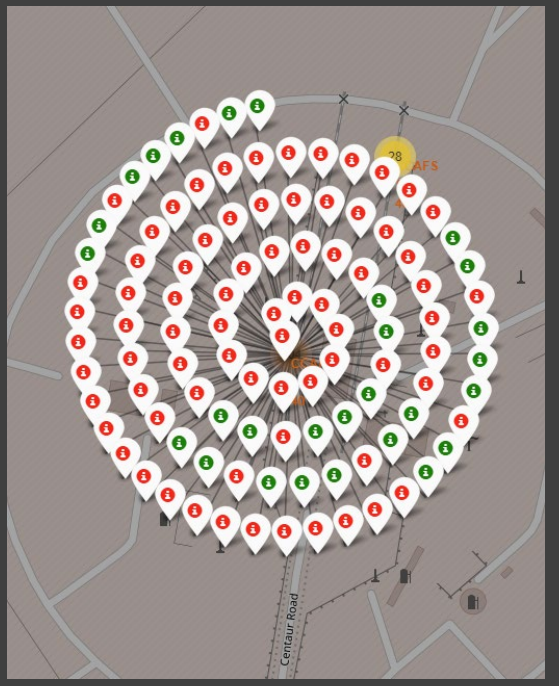
KSC LC-39A



CCAFS SLC-40



VAFB SLC-4E

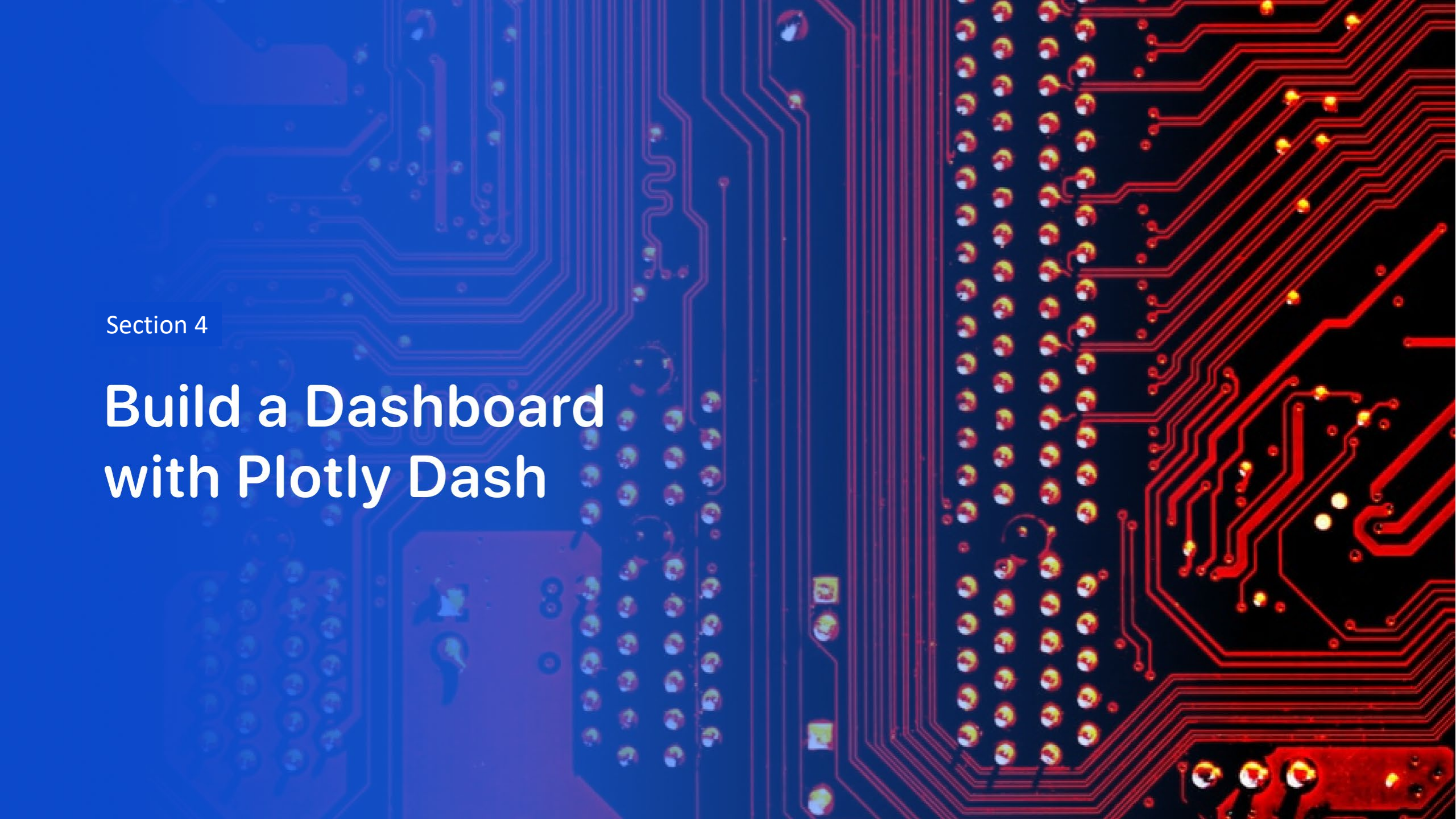


CCAFS LC-40

# Success ratio at all sites

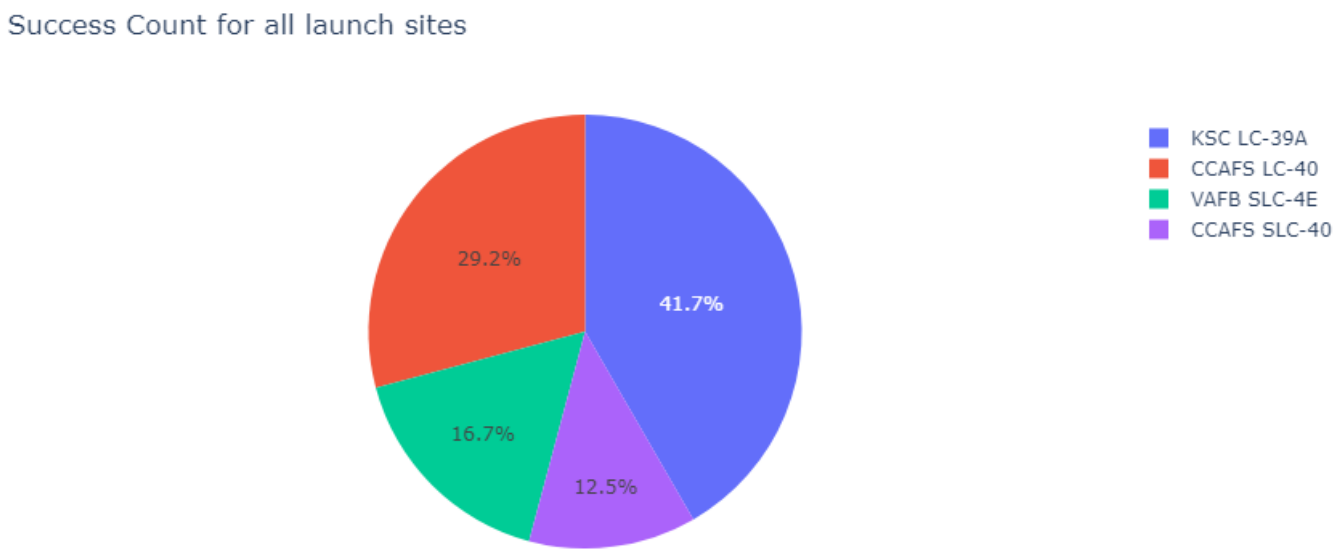
The green markers represent Success, and the red markers represent Failure.

While “CCAFS LC-40” demonstrates the lowest success rate, it clearly has the highest number of lunches. “KSC LC-39A” has the highest success rate by having a decent number of launches.


Section 4
Build a Dashboard with Plotly Dash

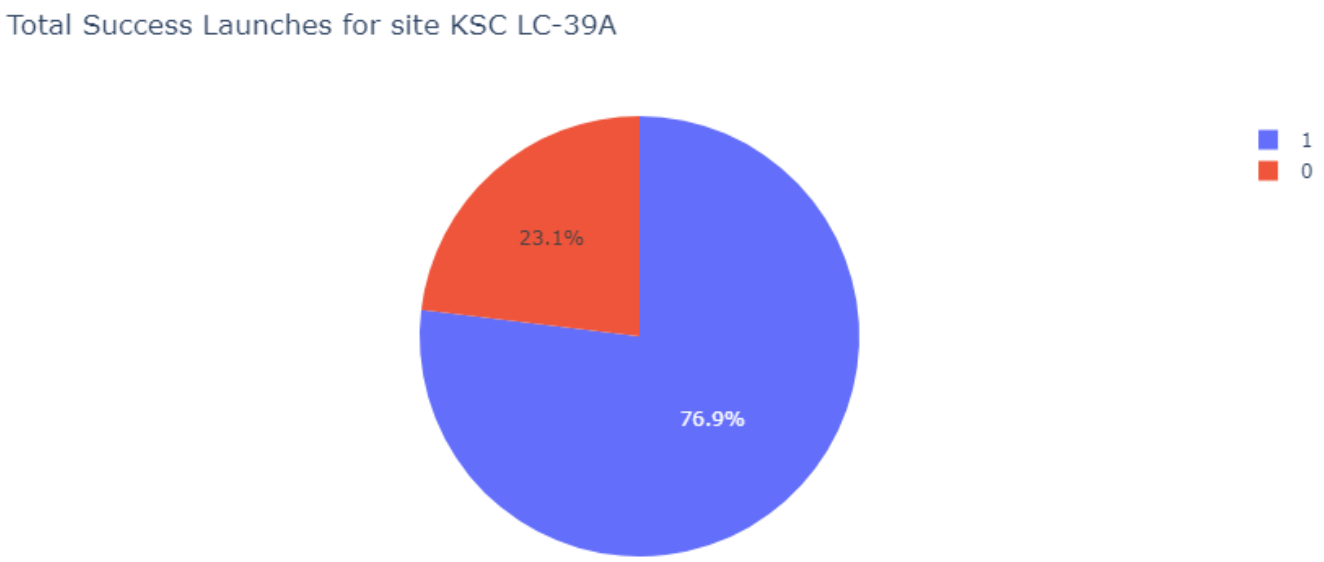
# Success count for all launch sites

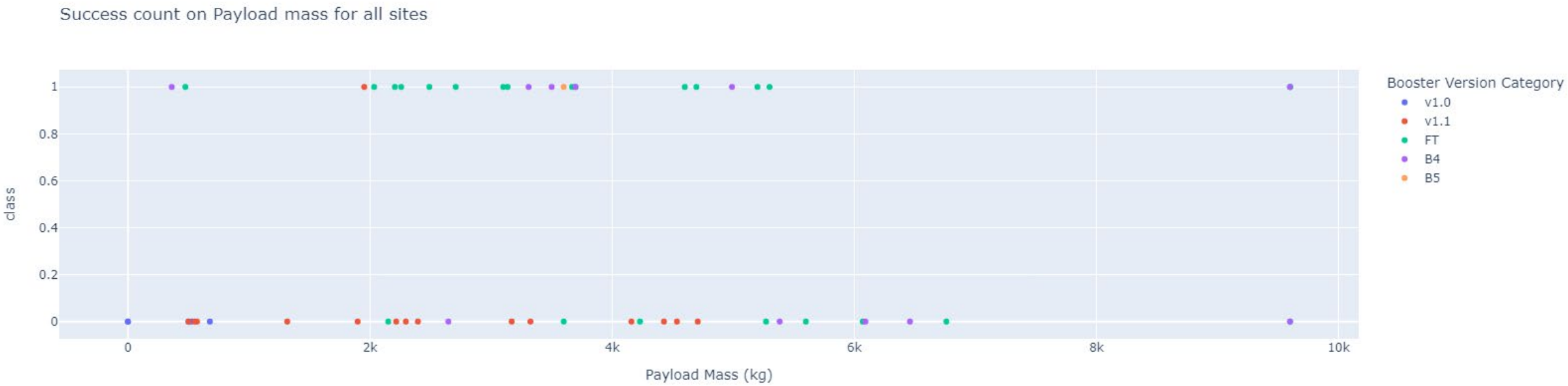
The Success rate at “KSC LC-39A” is significantly higher than other launch sites.



# Success/Failure ratio at “KSC LC-39A”

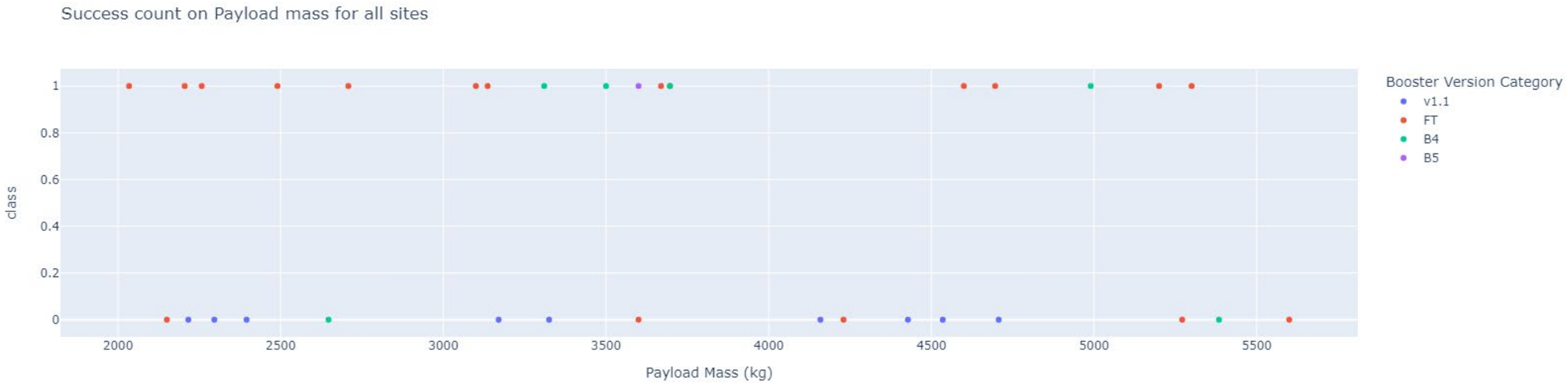
The success rate at KSC LC-39A is at approximately 77% which deems to be considerable!






# Success count on Payload mass for all sites

This plot shows the payload mass between zero and 10,000kg. Overall, “FT” followed by “B4” booster versions seem to have higher number of successes. We also can clearly see that payload mass for many launches fall between 2000Kg and 6000kg.




# Success count on Payload mass (2k – 6k) for all sites

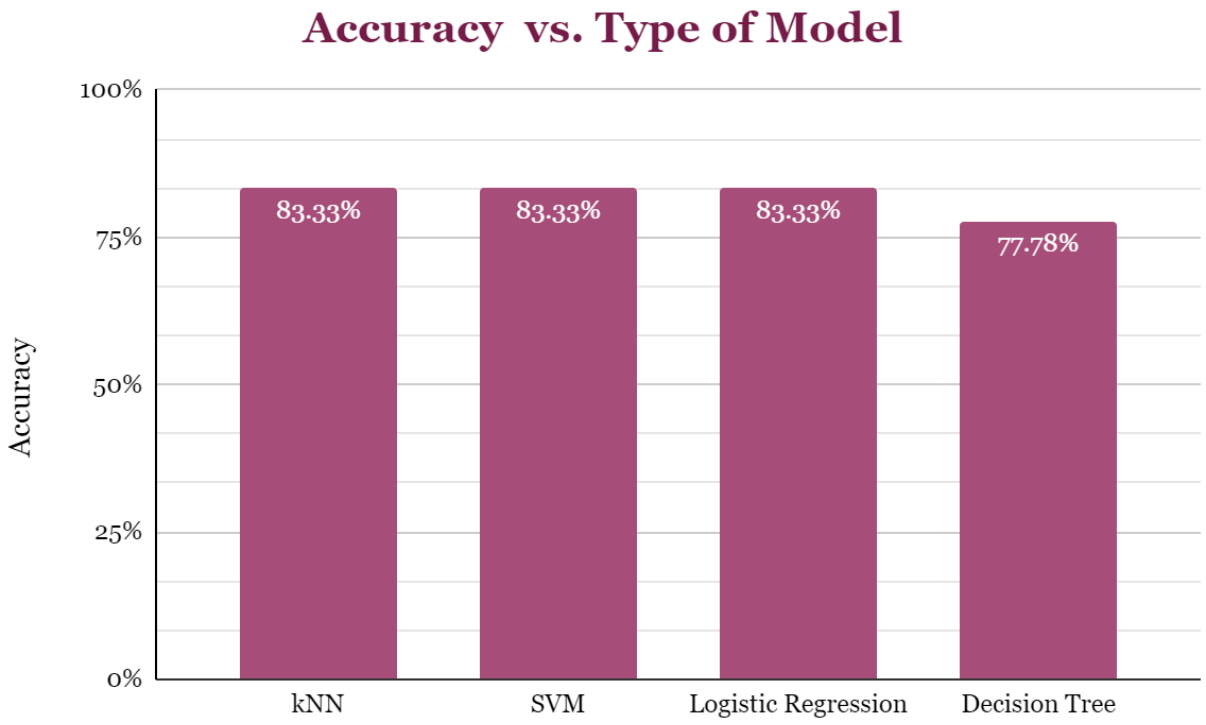
This plot shows the payload mass between 2000Kg and 6000kg.
In this range “FT” also holds the highest number of successes.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

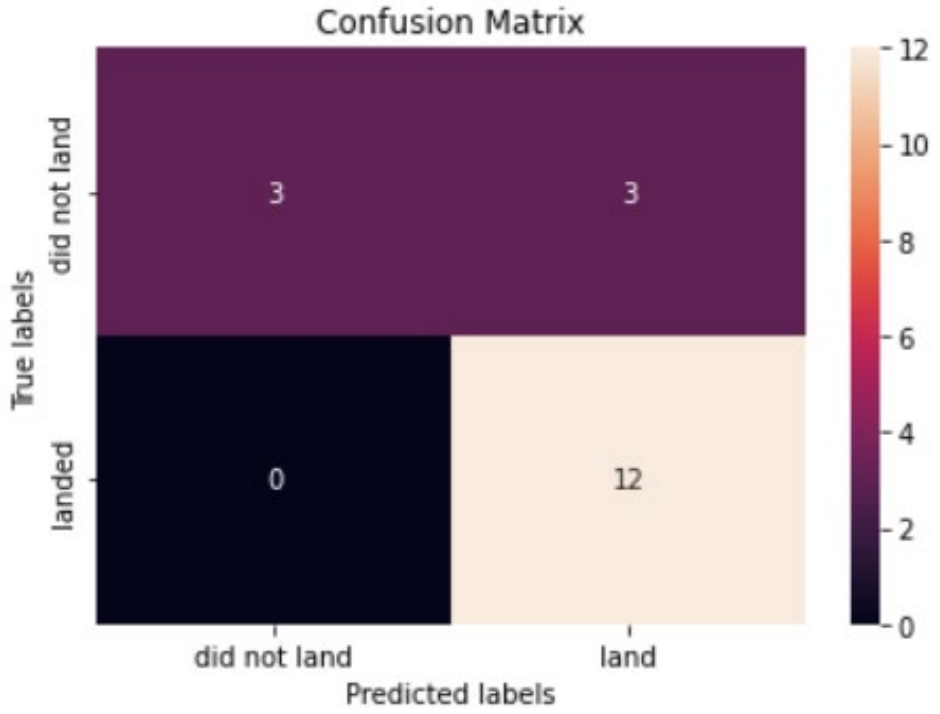
- kNN, SVM and Logistic Regression models all have an equal accuracy of 83.3% which is also is the highest accuracy achieved.
- The accuracy of the Decision Tree model, is slightly lower at 77.8%.

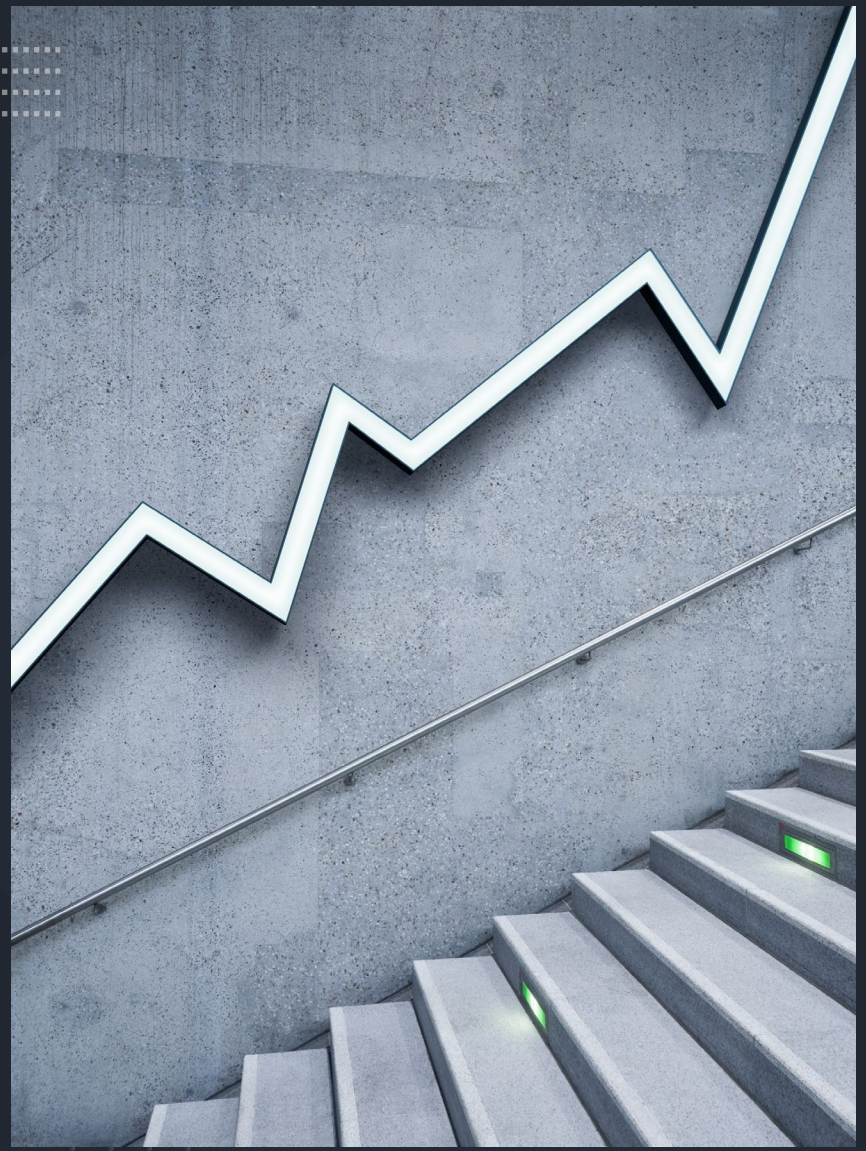


# Confusion Matrix

The confusion Matrix for all the three high performed models were identical as shown in this chart.

The models only predicted 3 out of 18 instances incorrectly!





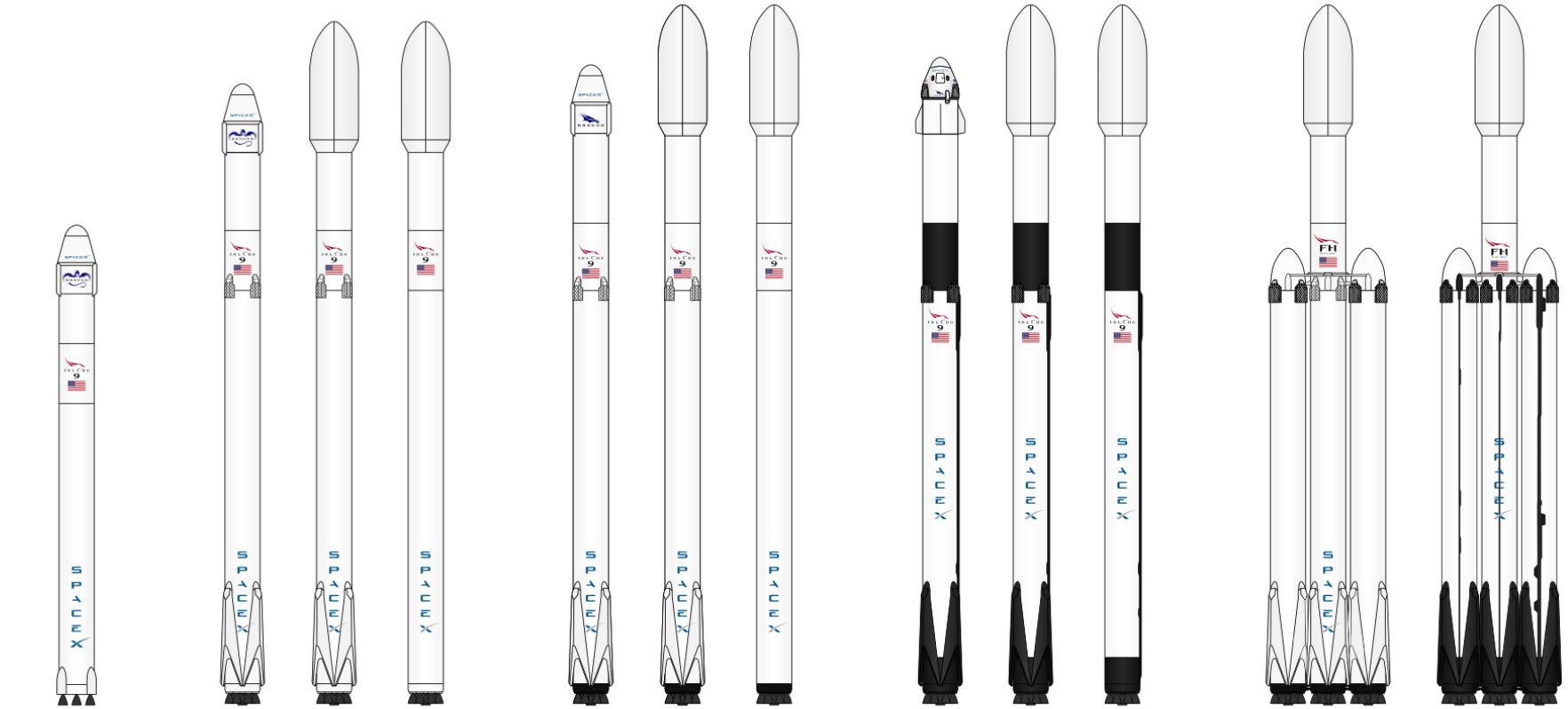
# Conclusions

- The success rate of stage 1 landing outcome has an overall upward trend since 2013. Therefore, the strategy that SpaceX has been using can be considered for our business model.
- The success rate at KSC LC-39A is the highest and at approximately 77% which provides high profitability!
- In choosing the location of launch sites proximity to airports, railways, highways and coastlines are important for fast and flexible access to transportation. To avoid any safety concerns, large cities and crowded areas should be avoided.
- Higher payload mass provide higher success rate in general. Our Aerospace Engineers may be able to provide an explanation for this observation.
- A list of versions of boosters that have carried the Maximum payload of 15600Kg since 2010 have been extracted and can be used.
- We have 3 machine learning model which can predict the first stage landing outcome with a high accuracy of 83%. We can use this to our advantage to optimize the outcome prior to launching the rockets!!!


47

# References


- *IBM Data Science Professional Certificate via Coursera: https://www.coursera.org/professional-certificates/ibm-data-science#courses*
- *Wikipedia - List of Falcon 9 and Falcon Heavy launches: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches*




List of Falcon 9 and Falcon Heavy launches

# Appendix

# Thank you!