Winning the Space Race with Data Science

Nicholas Chew
6th December 2023

Outline

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

Executive Summary

Summary of methodologies

- Data collection through API
- Data Collection through Web Scraping
- Data Wrangling with Python
- Exploratory Data Analysis
- Create visualization analytics dashboard with Folium and Plotly Dash
- Predictive Analysis – test multi models if the first stage of F9 will land succesfully

Summary of all results

Using K-nearest neighbour with one of the best accuracy models, we can predict that only 41% of all first stage flight will successfully land

Introduction

Project background

In this project, we want to predict if the Falcon 9 first stage will land successfully (problem).

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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

Methodology

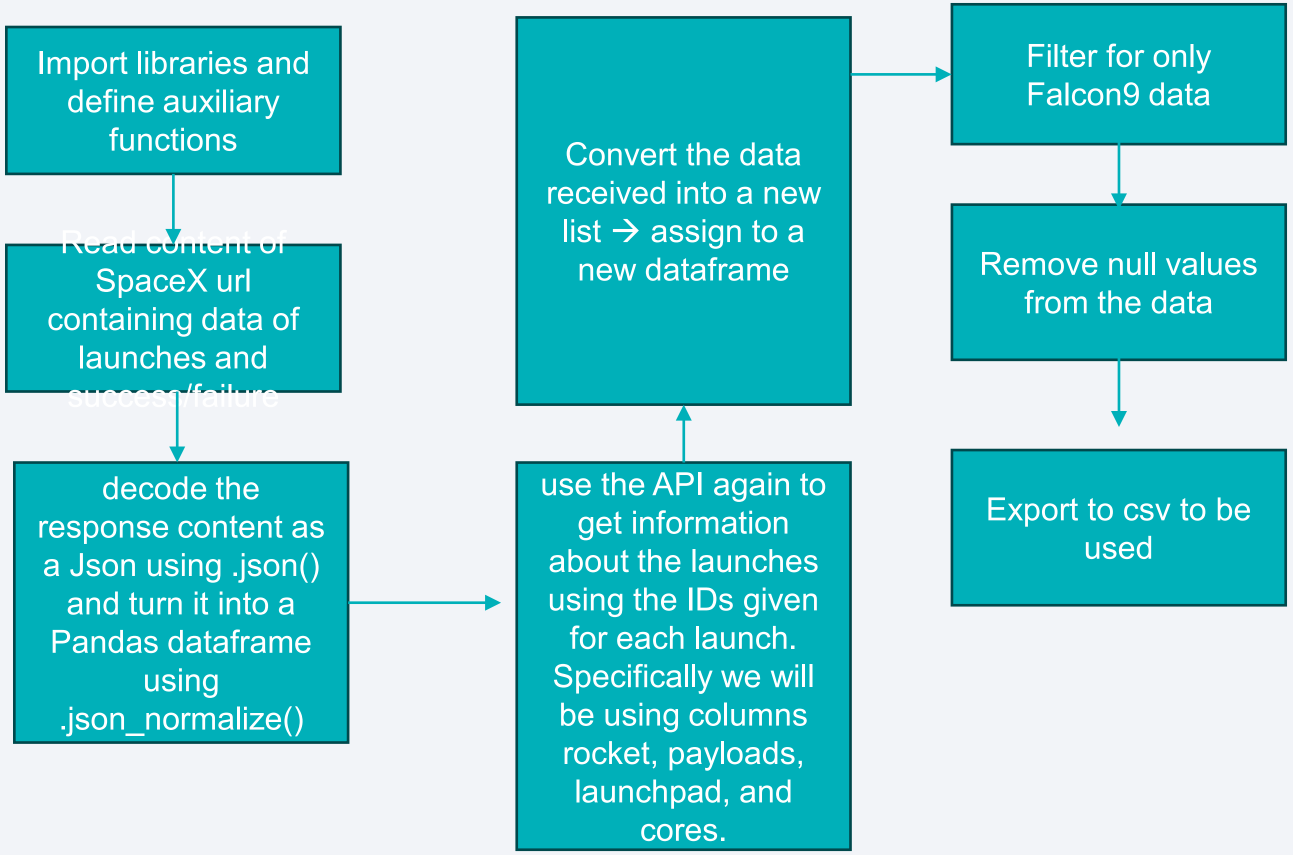
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through API from SpaceX's site and Wikipedia through web scrapping
- Perform data wrangling
  - Data was grouped through primary category including 'Class', 'Booster Version', 'Orbit', and 'Launch Site' with a count of its success/failures.
  - Any columns with null values were dropped.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - In the predictive analysis using classification models, we employed several methodologies to evaluate and select the best-performing model. We utilized techniques such as cross-validation, hyperparameter tuning, and careful experimentation with different machine learning methods to assess their performance and identify the most effective approach for the given task. This involved creating various classification models, such as logistic regression, support vector machines, decision trees, and k-nearest neighbors, and then using GridSearchCV to find the best parameters for each model. By leveraging these methodologies, we were able to determine the most suitable model for the predictive analysis, ensuring robust performance and reliable predictions for the given problem.

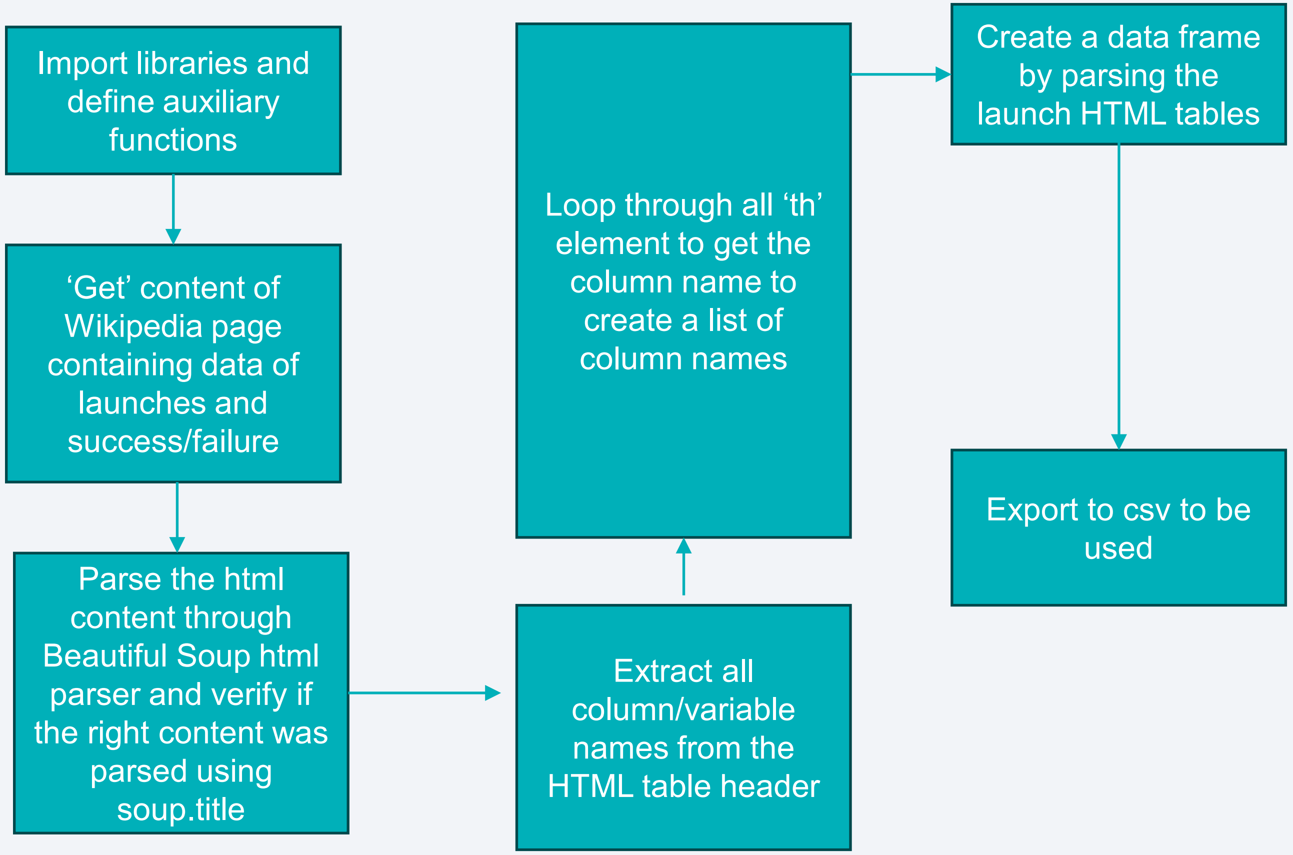
Data Collection – SpaceX API

- There were two approaches in data collection with the first being an API call to https://api.spacexdata.com/v4/launches/past.
- This link contains spaceX launches data, including rocket specifications and its success. The following flow chart on the right shows the process in which the API was called and data was cleaned.
- Github link to notebook



Data Collection - Scraping

- The second approach in data collection involves ‘scrapping’ data from SpaceX’s Wikipedia page in the form of html code.
- This approach uses the ‘BeautifulSoup’ library.
- This link contains spaceX launches data, including rocket specifications and its success. The following flow chart on the right shows the process in which the API was called and data was cleaned.
- Github link to notebook



Data Wrangling

- In data wrangling, we prepare the data through exploring the data before transforming it into a workable flat file.
- Github link to notebook

Define auxiliary functions

Read the csv prepared from the earlier data collection process

Identify data types of each columns

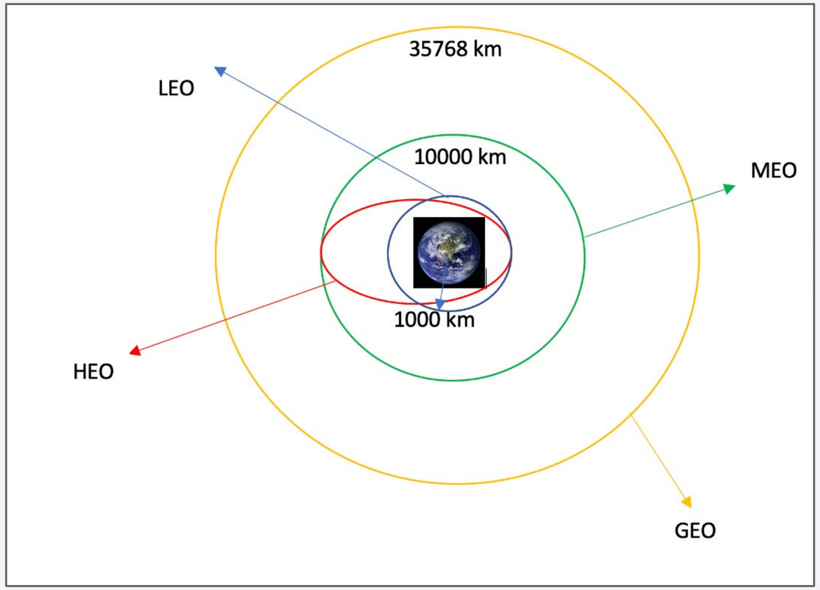
Calculate number of launches on each site

Calculate number and occurrence of each orbit

Calculate number and occurrence of mission outcomes of the orbit

Creating a landing outcome label from the Outcome column

Export data to .csv to be used



EDA with Data Visualization

- Chart 1: A categorical chart was plotted to explore the relationship between flight number, launch site, and mission success
- Chart 2: A categorical chart was plotted to explore the relationship between pay load mass (kg), launch site, and mission success
- Chart 3: A bar chart was plotted to explore the relationship between orbit and success rate
- Chart 4: A categorical chart was plotted to explore the relationship between orbit, flight number, and mission success
- Chart 5: A categorical chart was plotted to explore the relationship between Orbit, pay load mass (kg), and mission success
- Chart 6: A line chart was plotted to explore the relationship between success rate and year (time series).
- GitHub link to notebook

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. Use a subquery
- 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 notebook link

Build an Interactive Map with Folium

- Markers were added to launch sites on the global map to show the locations of launch sites.
- Markers within launch sites were added to indicate count of missions, and success/failure of missions. Green represents success missions, Red represents failed missions.
- Proximity lines were added to show the distance between launch sites and important locations/landmarks.

[Github notebook link]

Predictive Analysis (Classification)

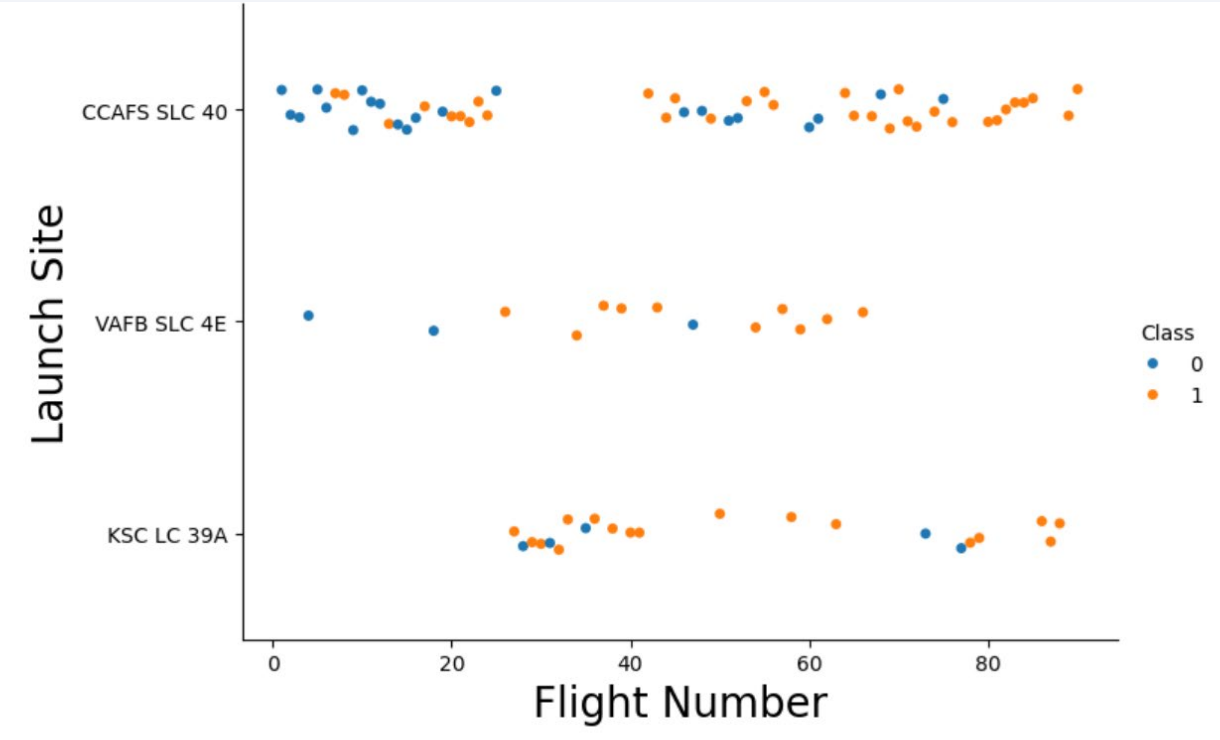
1. Data was split to test and train data
2. Create GridSearchCV object with cv=10 for optimization purposes
3. Apply GridSearchCV to different analysis methodology, including logistic regression, SVC, Decision Tree, and KNN
4. Calculate accuracy on test data
5. Assess confusion matrix for all outcome

Github notebook link

Insights drawn from EDA

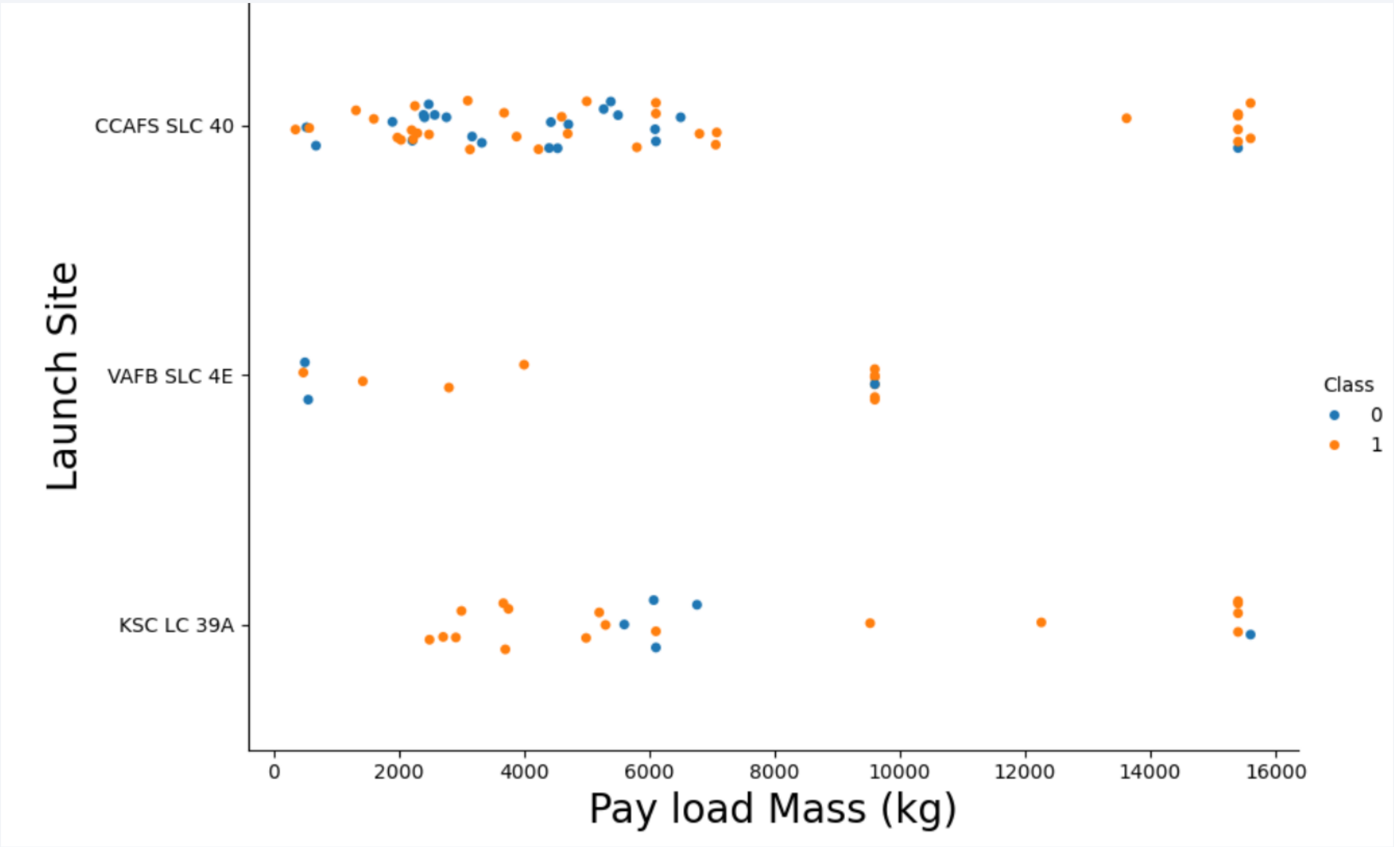
Flight Number vs. Launch Site

- In this chart, it shows that launch site looks to make a difference to mission success.
- Specifically, CCAFS SLC 40 has almost a 40-60 success rate, whilst the other two sites shows a higher success rate.



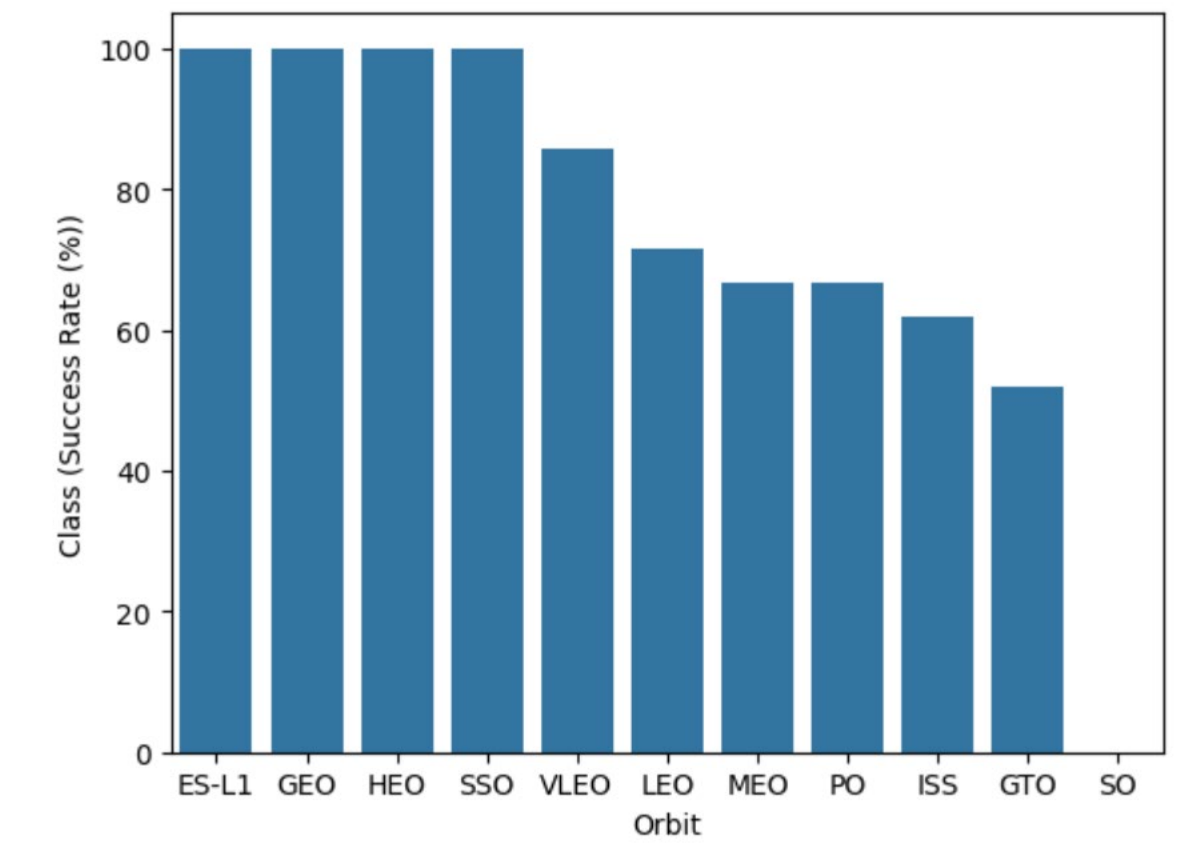
Payload vs. Launch Site

- At CCAFS SLC40 launch site, pay load mass doesn't look to have an impact to mission success. It looks to have an almost equal success/failure rate. It looks to have a higher success rate at 16000 kg payload though.
- At VAFB SLC 4E launch site, success rate looks to be higher between 1800 t0 4500 pay load
- At KSC LC 39A launch site, success rate looks higher below 6000 pay load mass, and failure was higher at around 6000 pay load.

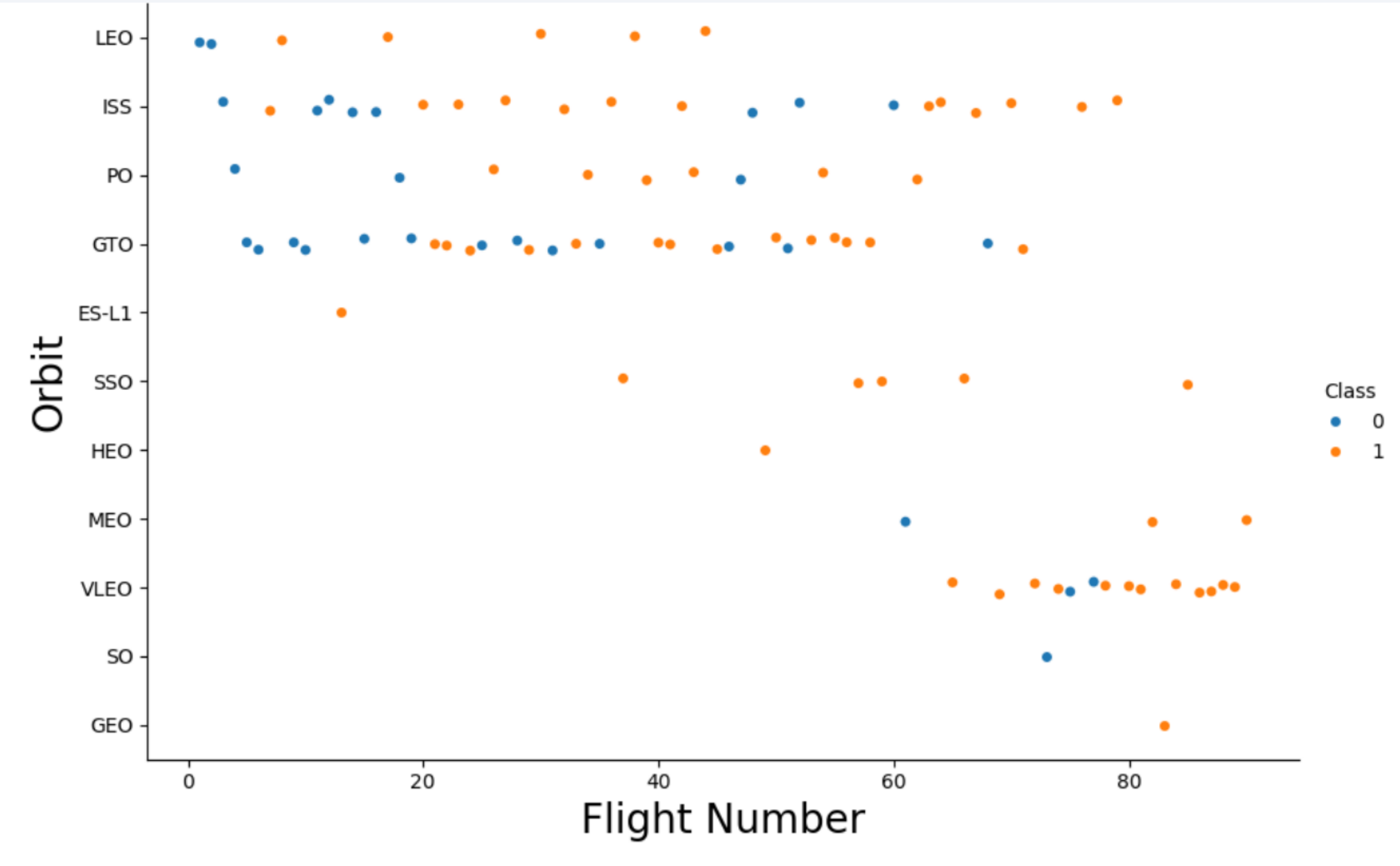


Success Rate vs. Orbit Type

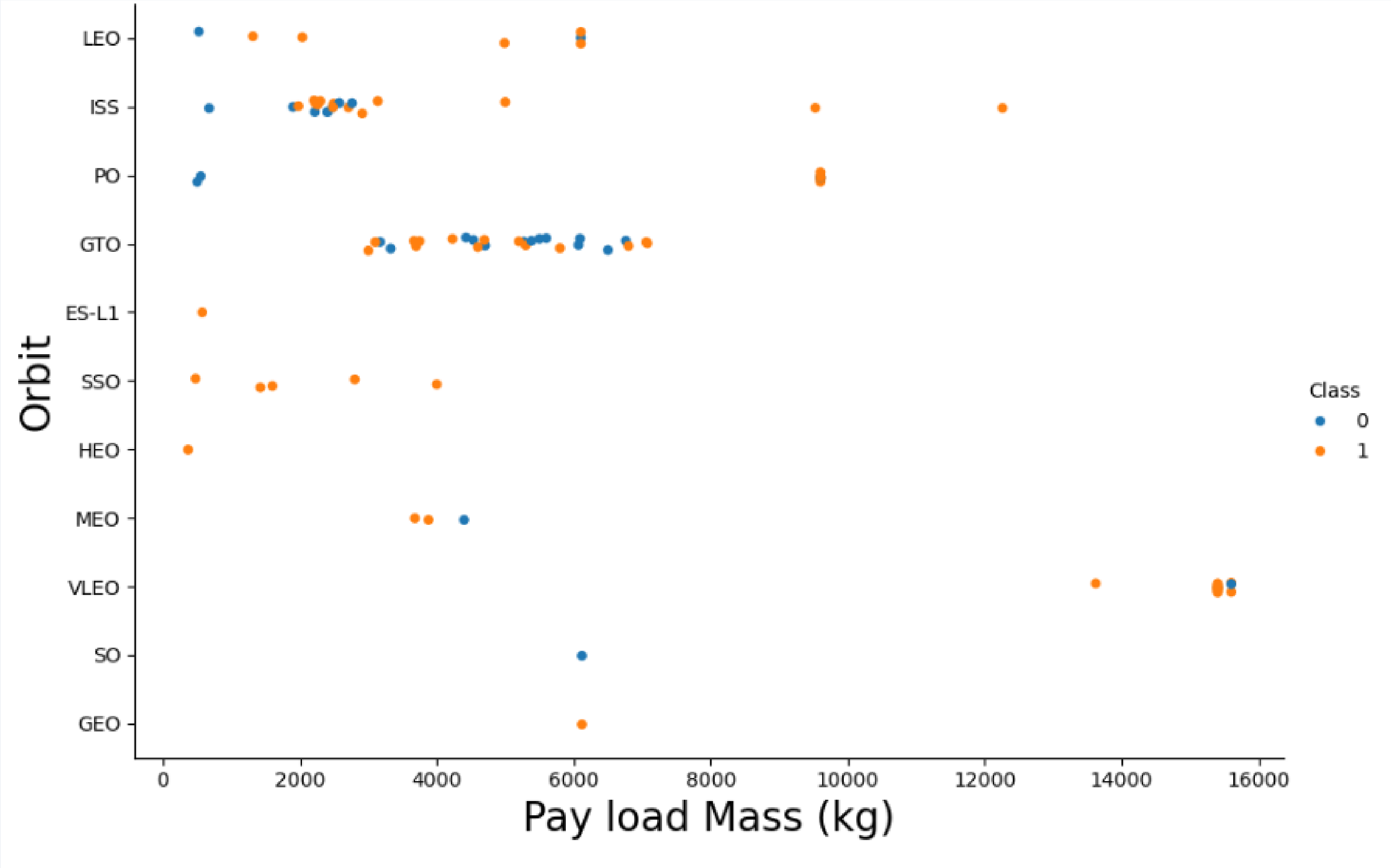
- The bar chart shows a notable differences in Orbit type vs success rate.
- ES-L1, GEO, HEO, and SSO had 100% success rate.
- Followed by VLEO at 83%.
- All other orbit types have <80% success rate.



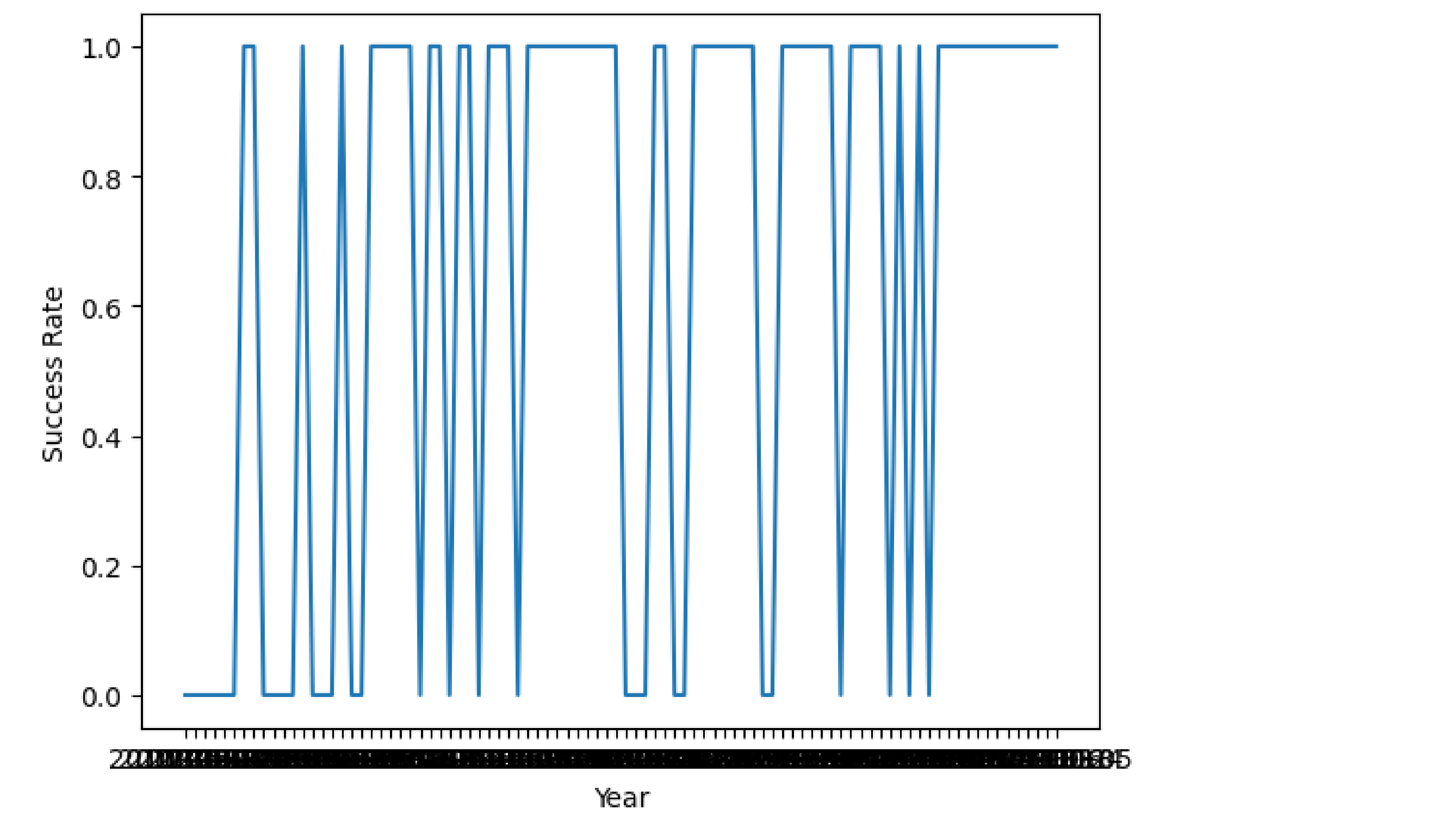
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

There were 4 unique launch sites in the dataset

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Following are 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lanc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failu
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	Failu
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

Total payload mass was 999,80kg

Average Payload Mass by F9 v1.1

The average payload mass by F9 v1.1 booster was 2534.67kg

First Successful Ground Landing Date

The first successful ground landing date was 22nd July 2018

Successful Drone Ship Landing with Payload between 4000 and 6000

Following are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

Following is a table showing the number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

Following are the names of the booster which have carried the maximum payload mass

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

2015 Launch Records

Following is a list of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Following is the rank of 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

Landing_Outcome	qty
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

Launch site proximity analysis

Launch Sites

This is a map of the United States of America pinpointing the launch sites of SpaceX
The pinpoints locations are marked in Yellow

Launch Outcomes

- In the launch site map below, green markers represent successful launches whilst red markers represent unsuccessful launches
- For the following launch site CCAFS SLC-40, it has a 3 out of 7 success rate

Distance to Proximities

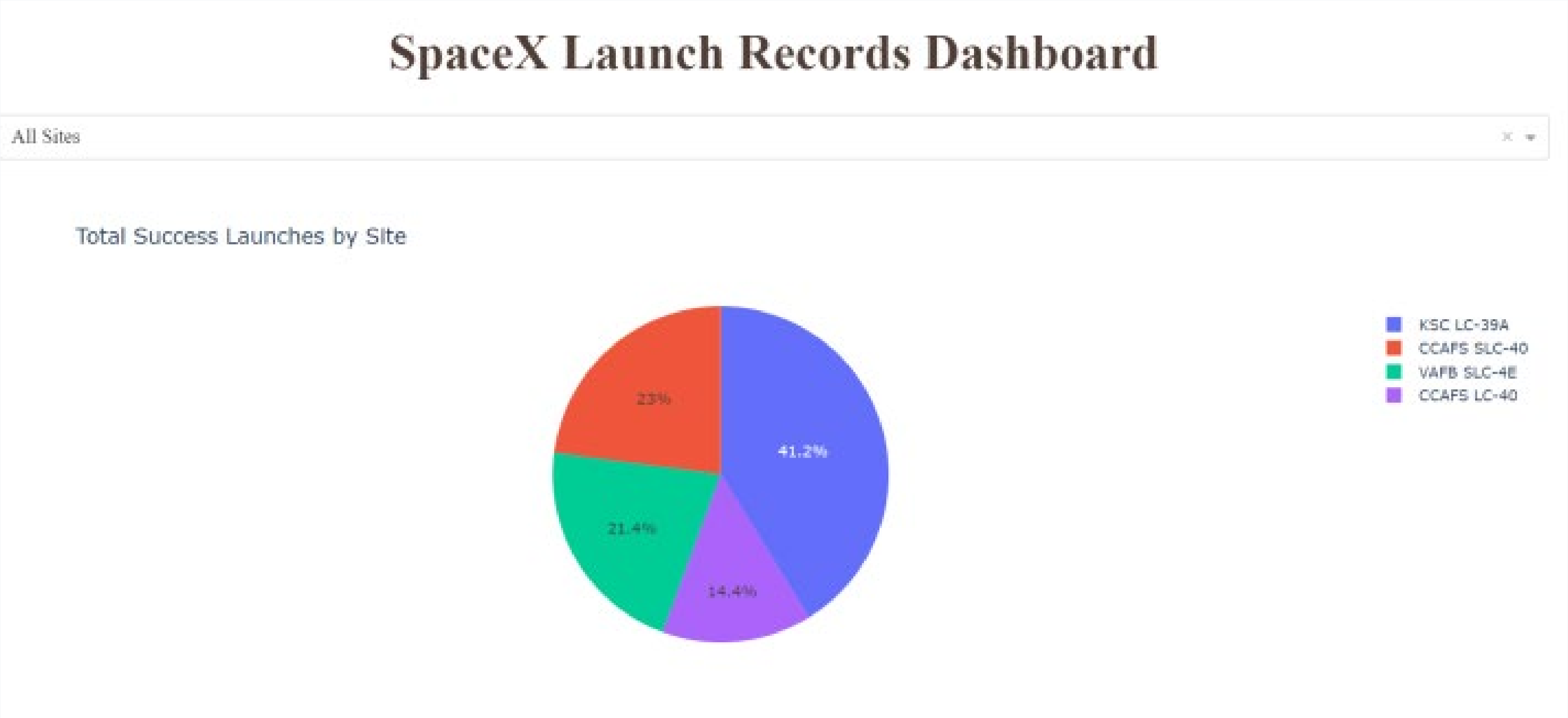
The launch site of CCAFS SLC-40 is:

- 23.23km from the nearest city
- 21.96km from the nearest railway
- 26.88 km from the nearest highway

Build a dashboard with Plotly Dash

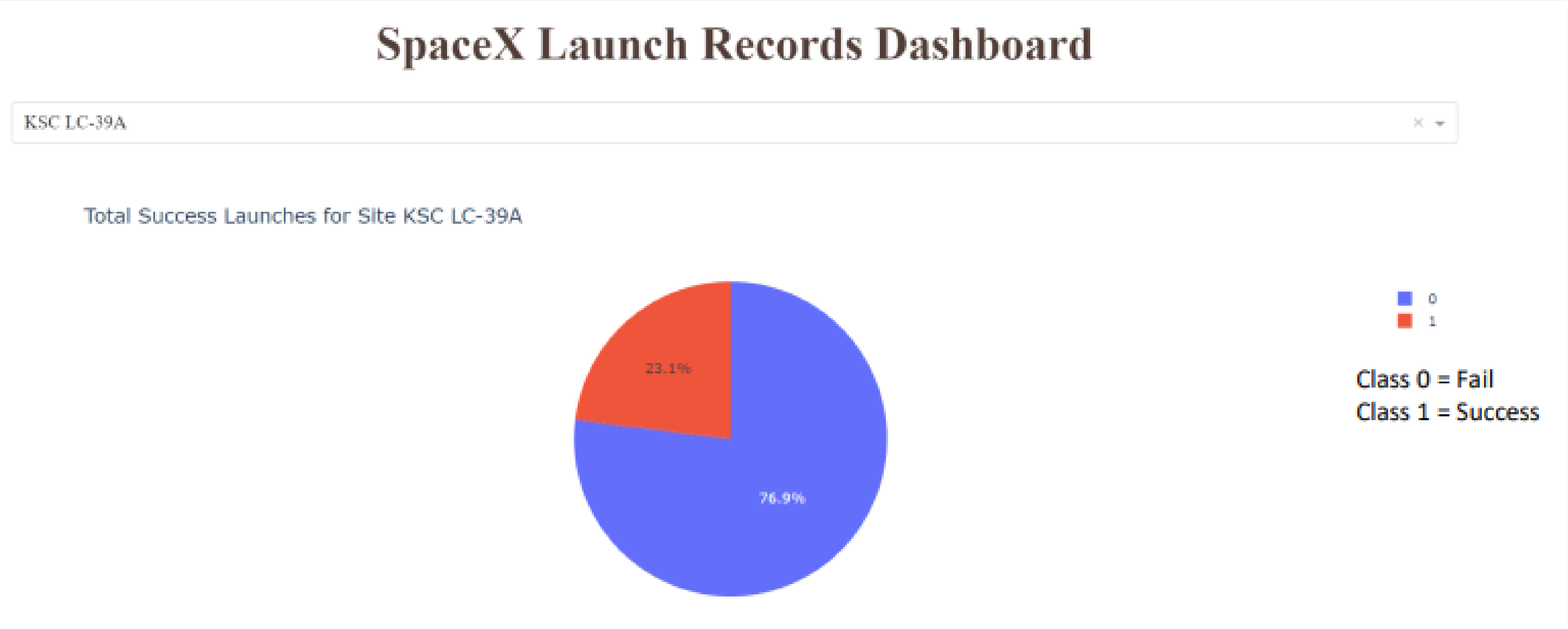
Total Launch Success by Site

Launch site KSC LC-39A has the highest success rate among other launch sites at 41.2%



Launch Success (KSC LC-29A)

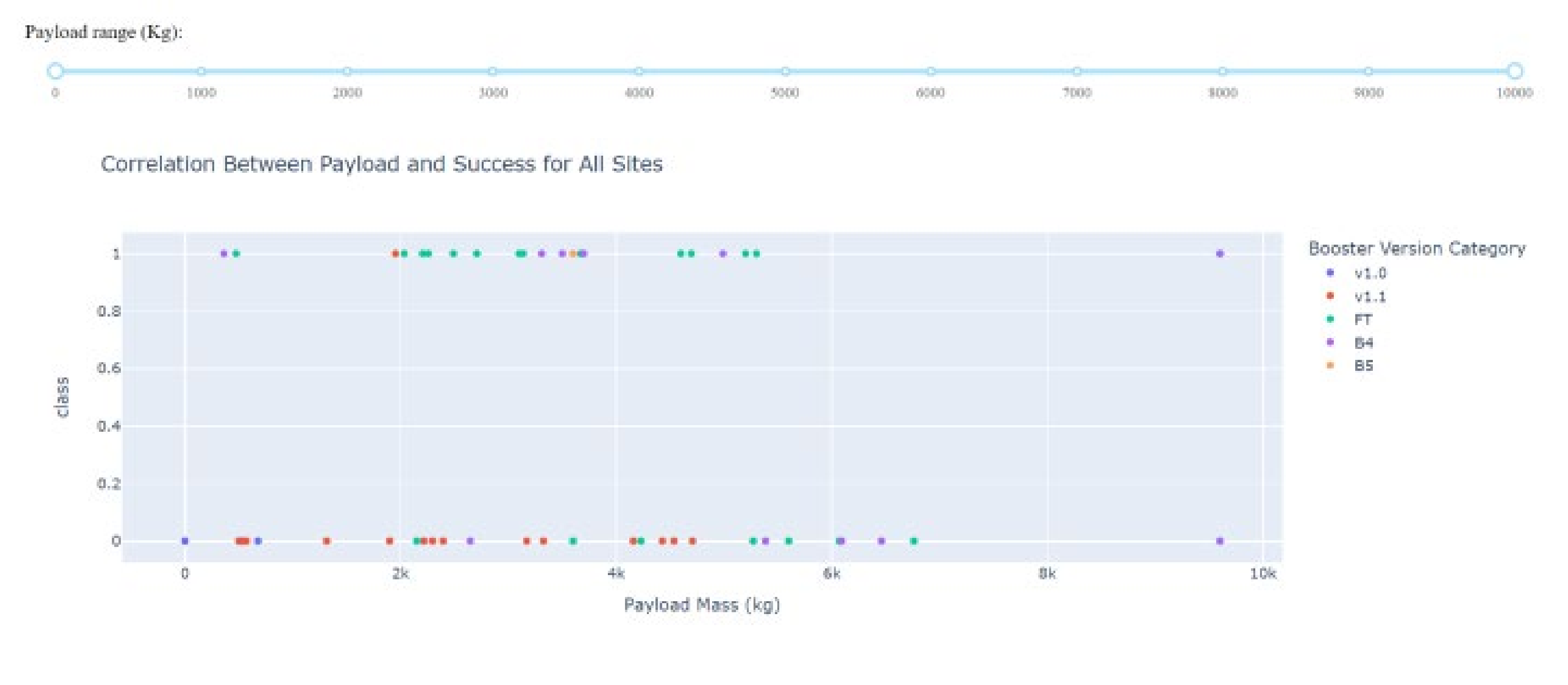
On its own launch site, its success rate was 76.9%



Payload Mass and Success

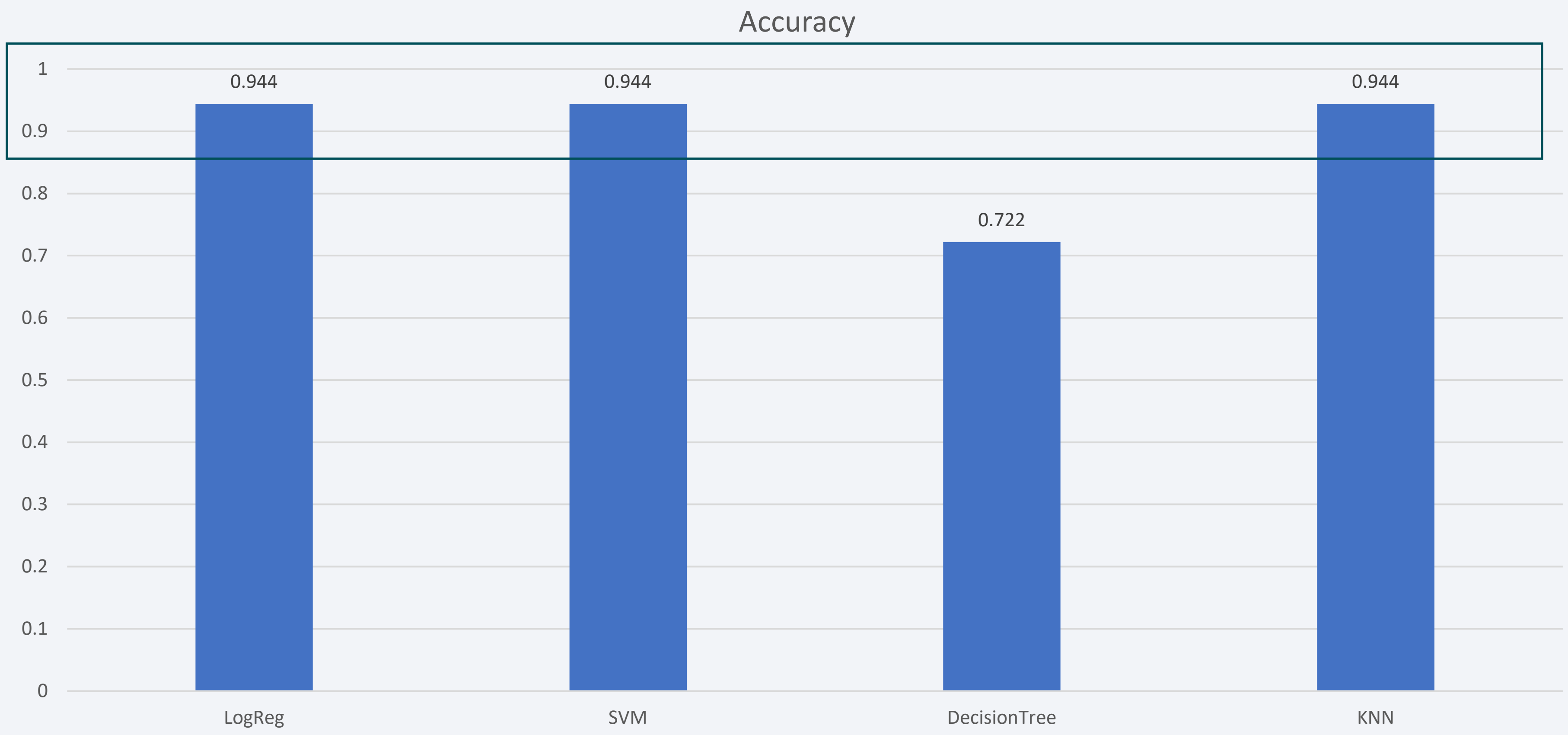
This dashboard shows a scatter plot exploring the relationship between payload mass and success rate.

The chart shows that payloads between 2,000kg and 5,000kg have the highest success rate.



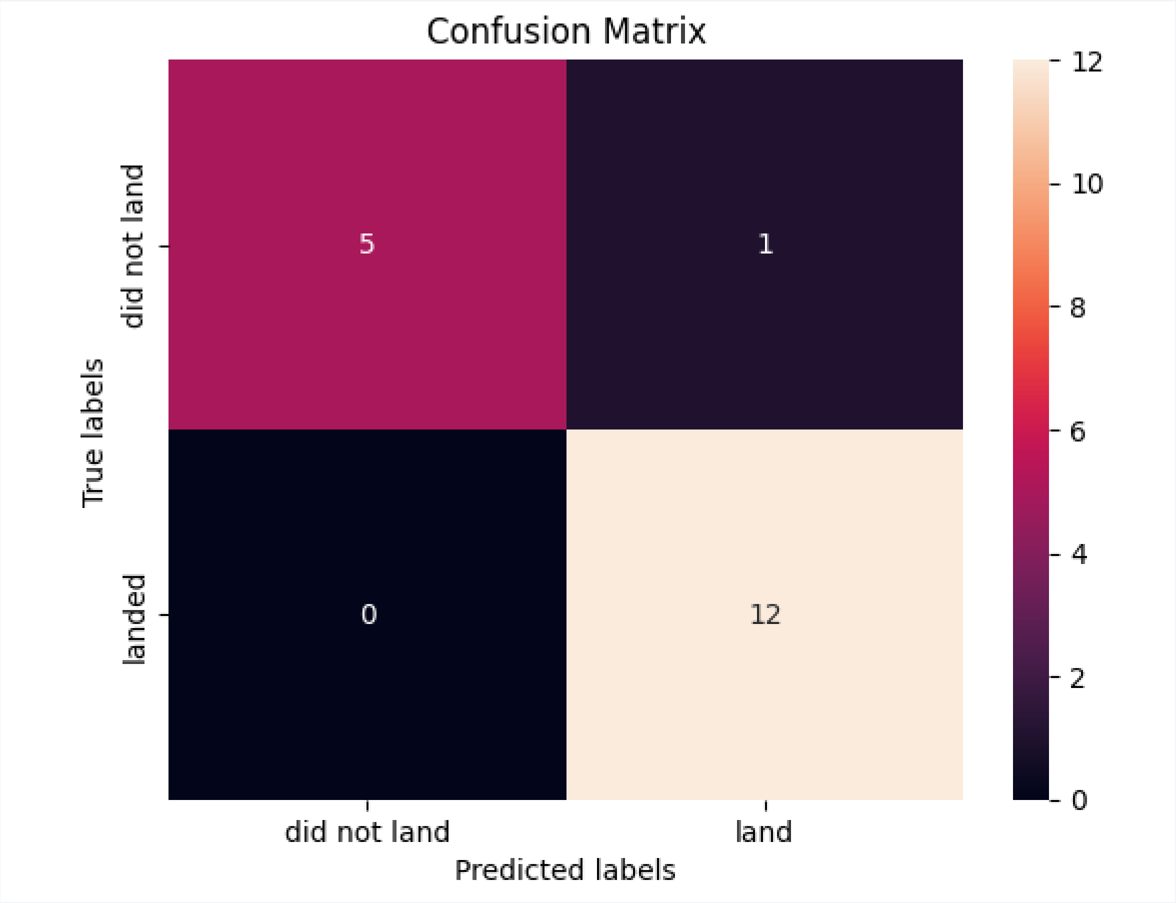
Predictive Analyses (Classification)

Classification Accuracy



Confusion Matrix

- This is a confusion matrix of K-nearest neighbour with test accuracy of 94.4% (cv = 10)



Conclusions

1. KSC LC-39A has the highest success rate among all launch sites with a 100% success rate for launches less than 5,000 kg.
2. Orbits ES-L1, HEO, SSO, and GEO have a 100% mission success rate.
3. On Payload Mass, the higher the payload mass, the higher the success rate.
4. Launch success increased over time. This shows that SpaceX learns from its mistake and constantly make correction to its rockets to increase its success rate.
5. The models in predictive analyses had similar accuracy, we can choose between Log Regression, Support Vector Machine, and KNN with all having a test accuracy of 94.4%

Thank You