Applied Data Science Capstone: SpaceX Falcon 9 launches

Mathias Wambeke

May 18th 2024

OUTLINE

- Executive Summary
- Introduction
- Methodology
  - Data collection
  - EDA & data wrangling
  - EDA & visual analytics
  - Predictive analysis
- Results
  - EDA with SQL
  - EDA with visualization
  - Map with folium
  - Dashboard
  - Predictive analysis
- Discussion: findings & implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- In this capstone, we will predict if the Falcon 9 first stage will land successfully
- This analysis can be applied for alternate companies bidding against SpaceX to launch rockets at a lower cost
- Our data is gathered from the SpaceX API, with additional information from publicly available websites
- The method is based on:
    - Exploratory Data Analysis (EDA) and Feature Engineering using Pandas, Matplotlib and Seaborn, together with SQL queries
    - Interactive visual analytics using Folium
    - Machine learning pipeline using Sklearn
- We find that successful landings are mainly determined by:
    - Flight number: increasing success rate over time
    - Payload: relatively high payloads are linked to a lower success rate
    - Orbit: lowest success rate for GTO orbit
- We conclude that a higher success rate is mainly achieved through experience over time and continuing improvements in technology

INTRODUCTION

- This presentation analyses successful landings of the first stage of the Falcon 9 rocket
- Having an accurate prediction of successful landings can have major implications:
  - SpaceX advertises Falcon 9 rocket launches at a cost of 62 million dollars
  - Other providers cost upward of 165 million dollars. Much of the savings from SpaceX are due to the reuse of the first stage.
  - If we can predict whether 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: DATA COLLECTION

- Rocket launch data is gathered from the SpaceX API:
  - https://api.spacexdata.com/v4/launches/past
  - The extracted features are: rocket, payloads, launchpad, cores, flight_number & date
  - We restrict the dates of the launches <= 2020/11/13
  - Booster version is restricted to Falcon 9 launches
  - PayloadMass missing values are replaced by the mean value of the dataset
- An additional data source is the Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches”:
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Web scraping is used to collect Falcon 9 historical launch records
  - The extracted features are: Flight No, Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date & Time

METHODOLOGY: EDA & DATA WRANGLING

- The data contains several Space X launch facilities:
  - Cape Canaveral Space Launch Complex 40 (CCAFS SLC 40)
  - Kennedy Space Center Launch Complex 39A (KSC LC 39A)
  - Vandenberg Air Force Base Space Launch Complex 4E (VAFB SLC 4E)
- The number of launches on each site is presented in the table below:

Launch Site	Count
CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

METHODOLOGY: EDA & DATA WRANGLING

- Each launch aims to an dedicated orbit:
  - Geostationary Transfer Orbit (GTO)
  - International Space Station (ISS)
  - Very Low Earth Orbits (VLEO)
  - Polar Orbit (PO)
  - Low Earth orbit (LEO)
  - Sun-Synchronous Orbit (SSO or SO)
  - Medium Earth Orbit (MEO)
  - Lagrange point 1 (ES-L1)
  - Highly Elliptical Orbit (HEO)
  - Geosynchronous Orbit (GEO)
- The number and occurrence of each orbit is presented in the table on the right:

Orbit	Count
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
ES-L1	1
HEO	1
SO	1
GEO	1

METHODOLOGY: EDA & DATA WRANGLING

- The type and number of landing outcomes is displayed in the table below
- The “Class” variable summarizes the mission outcome: if the value is zero, the first stage did not land successfully; one means the first stage landed successfully
- The overall success rate is 66,6%

Outcome	Mission outcome description	Count	Class
True ASDS	successfully landed to a drone ship	41	1
None None	failure to land	19	0
True RTLS	successfully landed to a ground pad	14	1
False ASDS	unsuccessfully landed to a drone ship	6	0
True Ocean	successfully landed to a specific region of the ocean	5	1
False Ocean	unsuccessfully landed to a specific region of the ocean	2	0
None ASDS	failure to land	2	0
False RTLS	unsuccessfully landed to a ground pad	1	0

METHODOLOGY: EDA & VISUAL ANALYTICS

- Exploratory Data Analysis (EDA) and Feature Engineering are performed using Pandas, Matplotlib and Seaborn
- We select the following features for success prediction: FlightNumber, PayloadMass, Orbit, LaunchSite, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial
- We apply one-hot encoding to create dummy variables from the categorical variables
- Interactive visual analytics are conducted using Folium

METHODOLOGY:PREDICTIVE ANALYSIS

- We create a machine learning pipeline to predict if the first stage will land
- Sklearn preprocessing is used to standardize our data
- Sklearn train_test_split is used to split into training data and test data
- Sklearn GridSearchCV is used to find best Hyperparameter for SVM, Classification Trees, Logistic Regression and K Nearest Neighbors
- We calculate the accuracy score to find the method that performs best using test data

RESULTS: EDA WITH SQL

- The results of the exploratory data analysis with SQL are presented in the table below:

total payload mass carried by boosters launched by NASA (CRS)	45596 (kg)
average payload mass carried by booster version F9 v1.1	2928.4 (kg)
date when the first successful landing outcome in ground pad was achieved	2015-12-22
boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000	F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
total number of successful and failure mission outcomes	101
names of the booster_versions 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

RESULTS: EDA WITH SQL

- Month, failure landing_outcomes in drone ship, launch_site & booster versions in 2015:

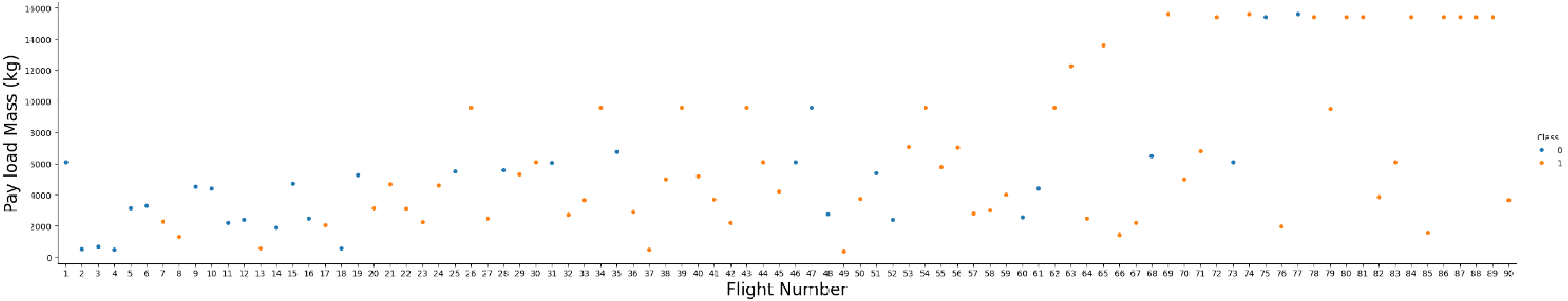
Month	Landing_Outcome	Launch_Site	Booster_Version
01	Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1012
04	Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1015

- Landing outcomes between the date 2010-06-04 and 2017-03-20:

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

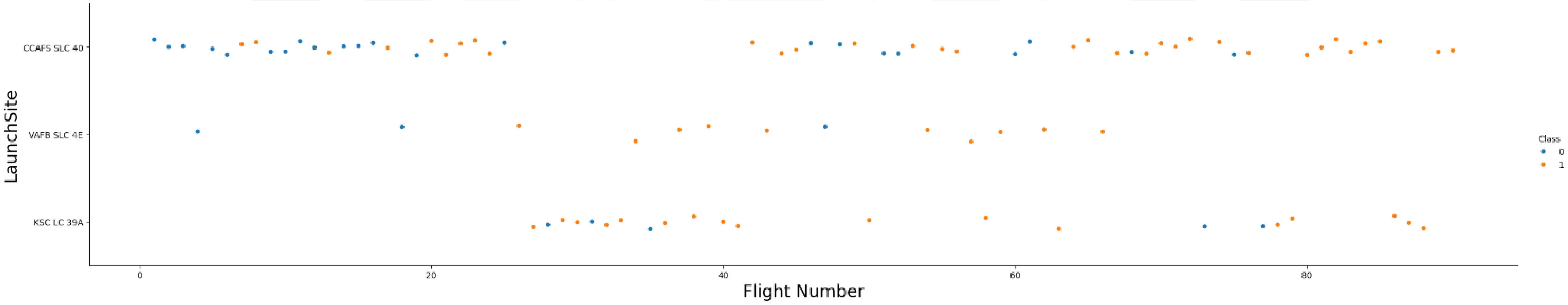
RESULTS: EDA WITH VISUALIZATION

- The figure below displays how the Flight Number (indicating the continuous launch attempts) and Payload variables would affect the launch outcome
- We see that as the flight number increases, the first stage is more likely to land successfully
- The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



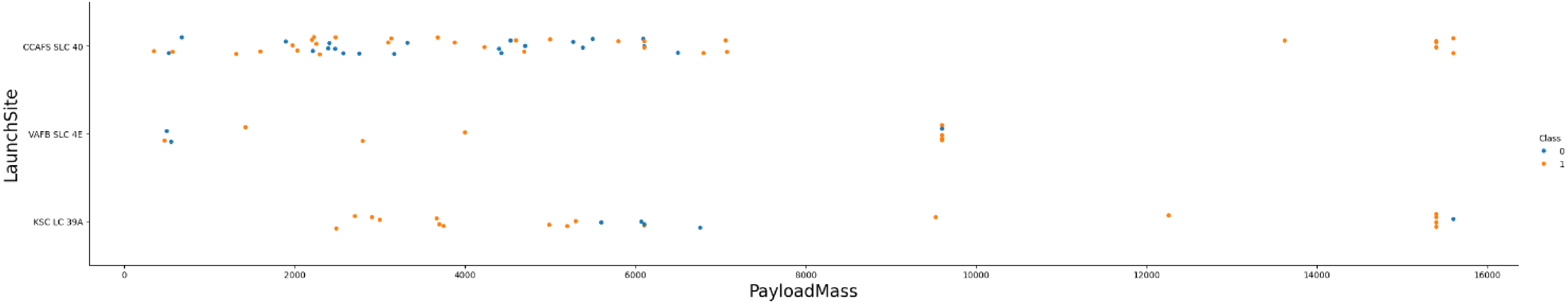
RESULTS: EDA WITH VISUALIZATION

- The figure below visualizes the relationship between Flight Number and Launch Site
- We see that different launch sites have different success rates. CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E have a success rate of 77%.



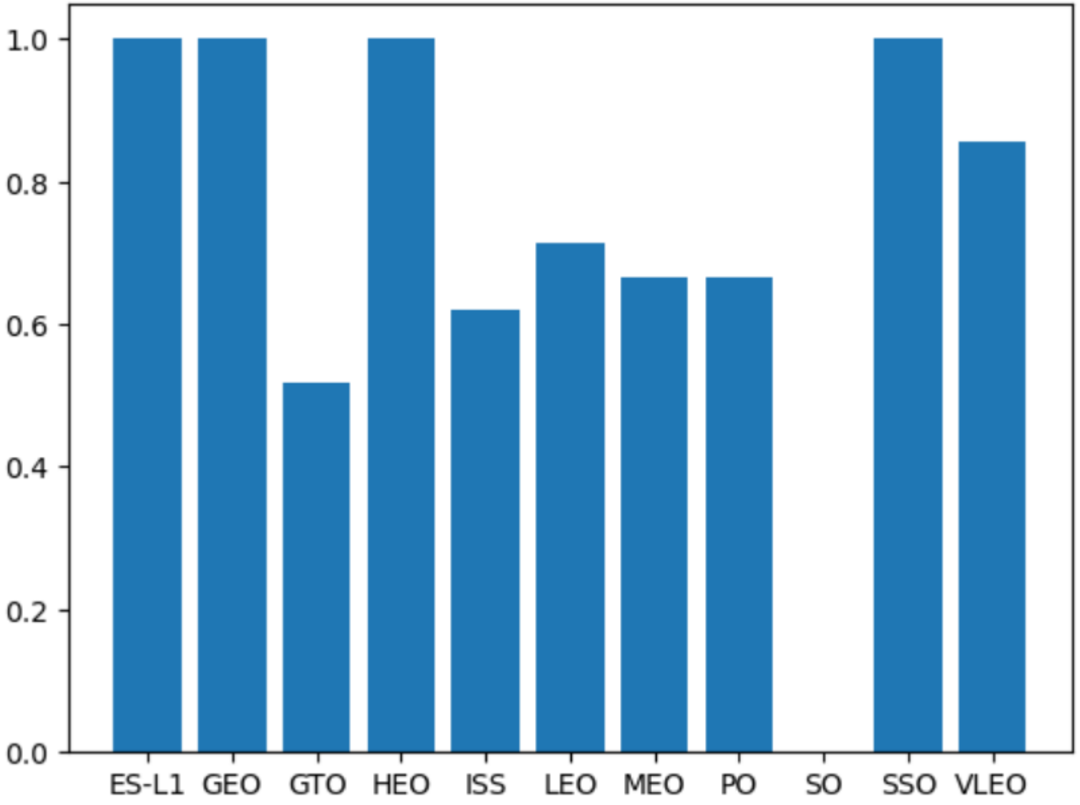
RESULTS: EDA WITH VISUALIZATION

- The figure below visualizes the relationship between Payload and Launch Site
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)



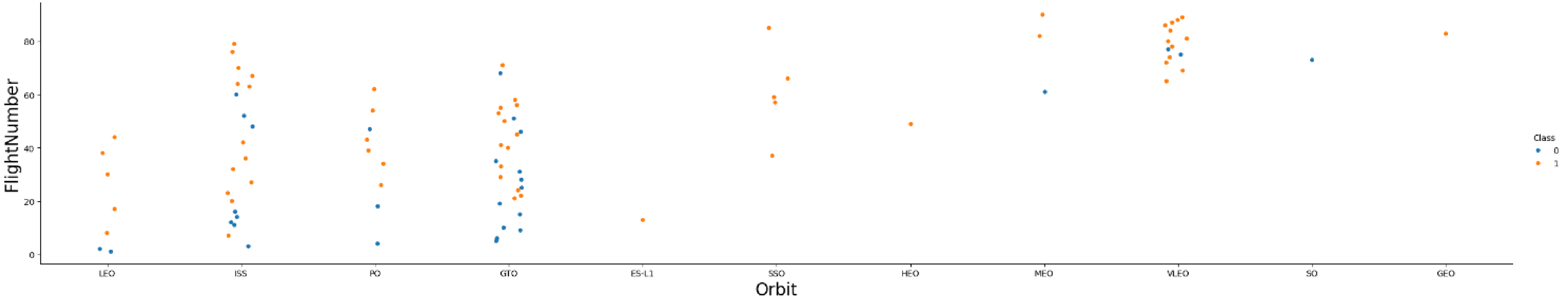
RESULTS: EDA WITH VISUALIZATION

- The figure below visualizes the relationship between the success rate of each orbit type



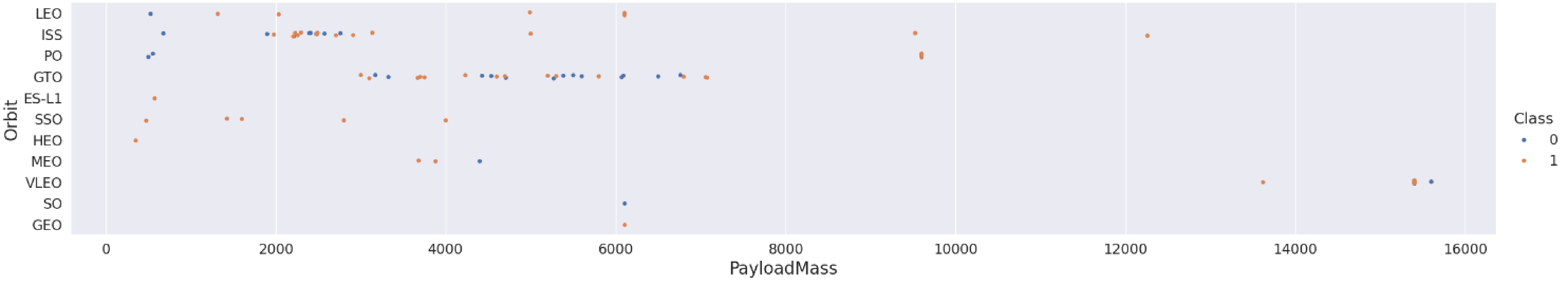
RESULTS: EDA WITH VISUALIZATION

- The figure below visualizes the relationship between Flight Number and Orbit type
- For 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.



RESULTS: EDA WITH VISUALIZATION

- The figure below visualizes the relationship between Payload and Orbit type
- With heavy payloads the successful landing are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive & negative landing rates are both observed for light & heavy payloads

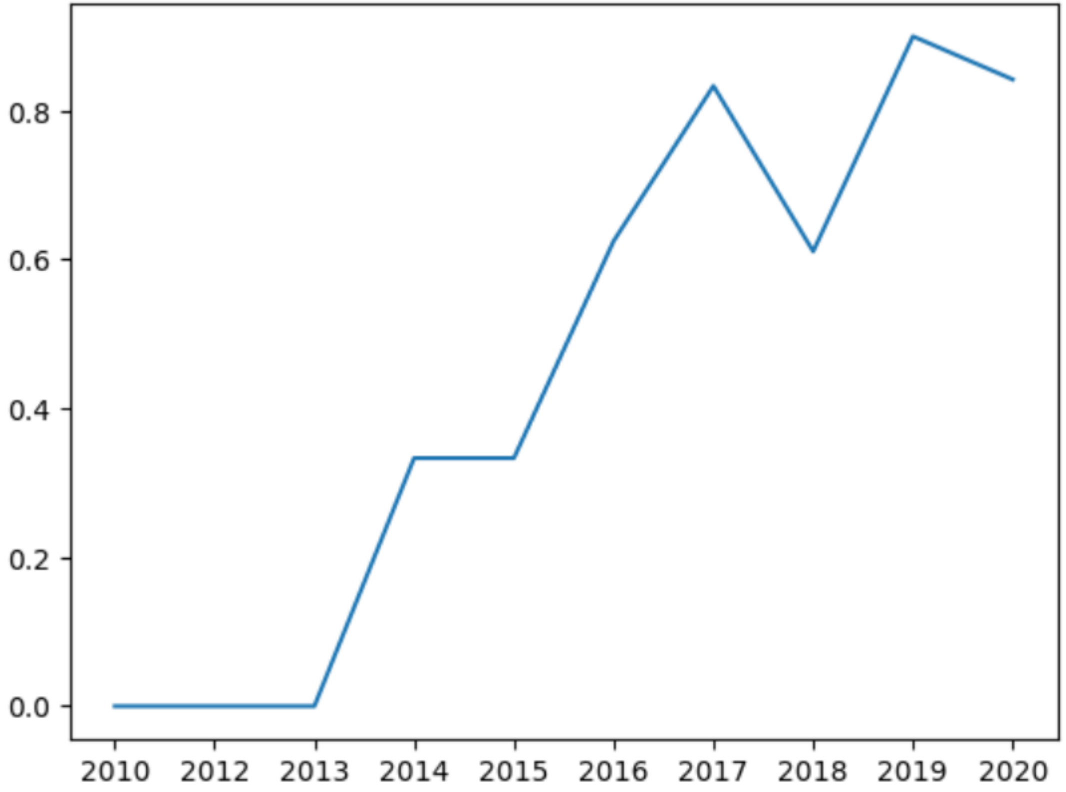


IBM Developer

SKILLS NETWORK

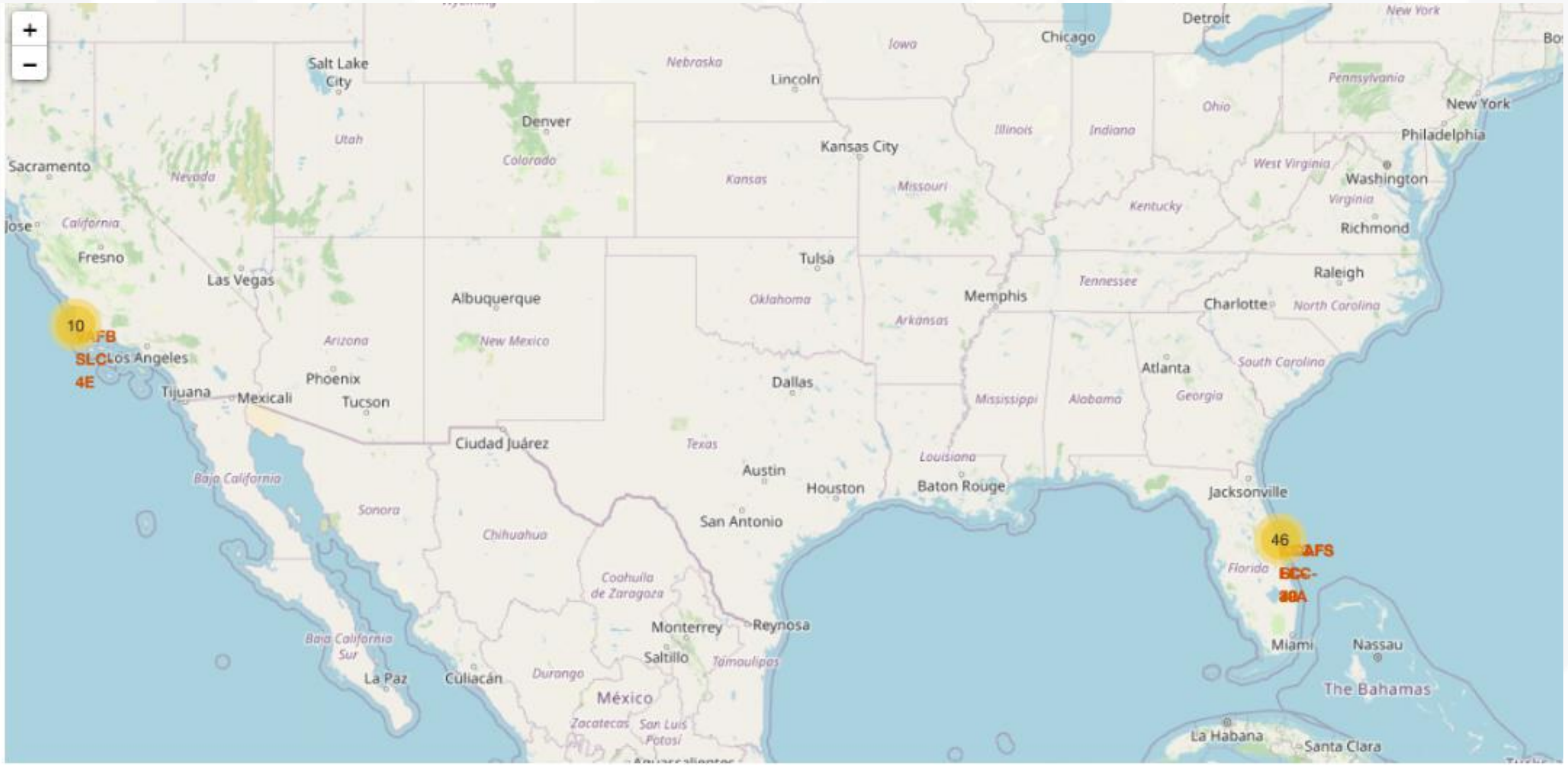
RESULTS: EDA WITH VISUALIZATION

- The figure below visualizes the launch success yearly trend
- The success rate since 2013 kept increasing till 2020



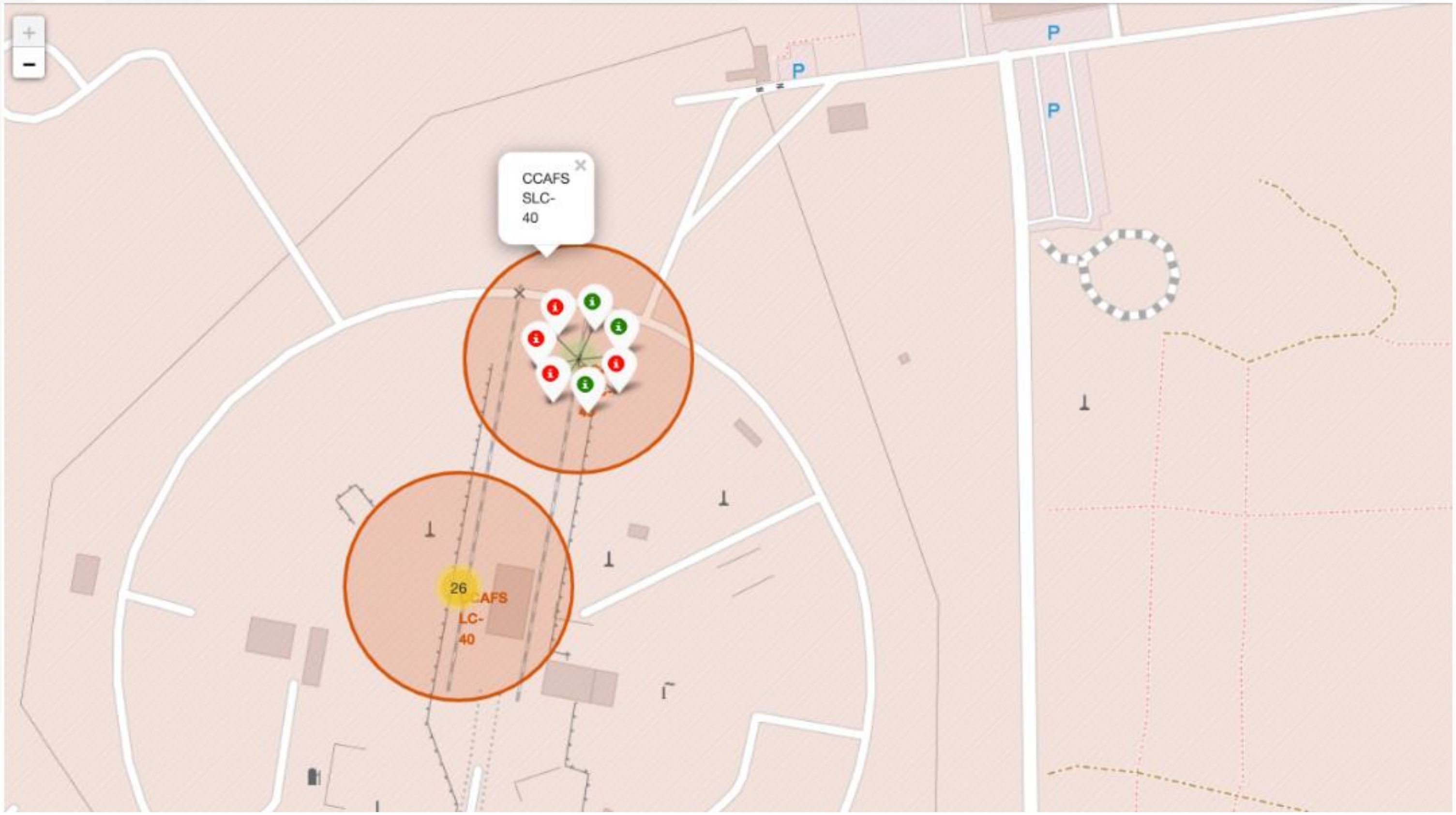
RESULTS: MAP WITH FOLIUM

- The figure below visualizes the launch sites and number of launches:



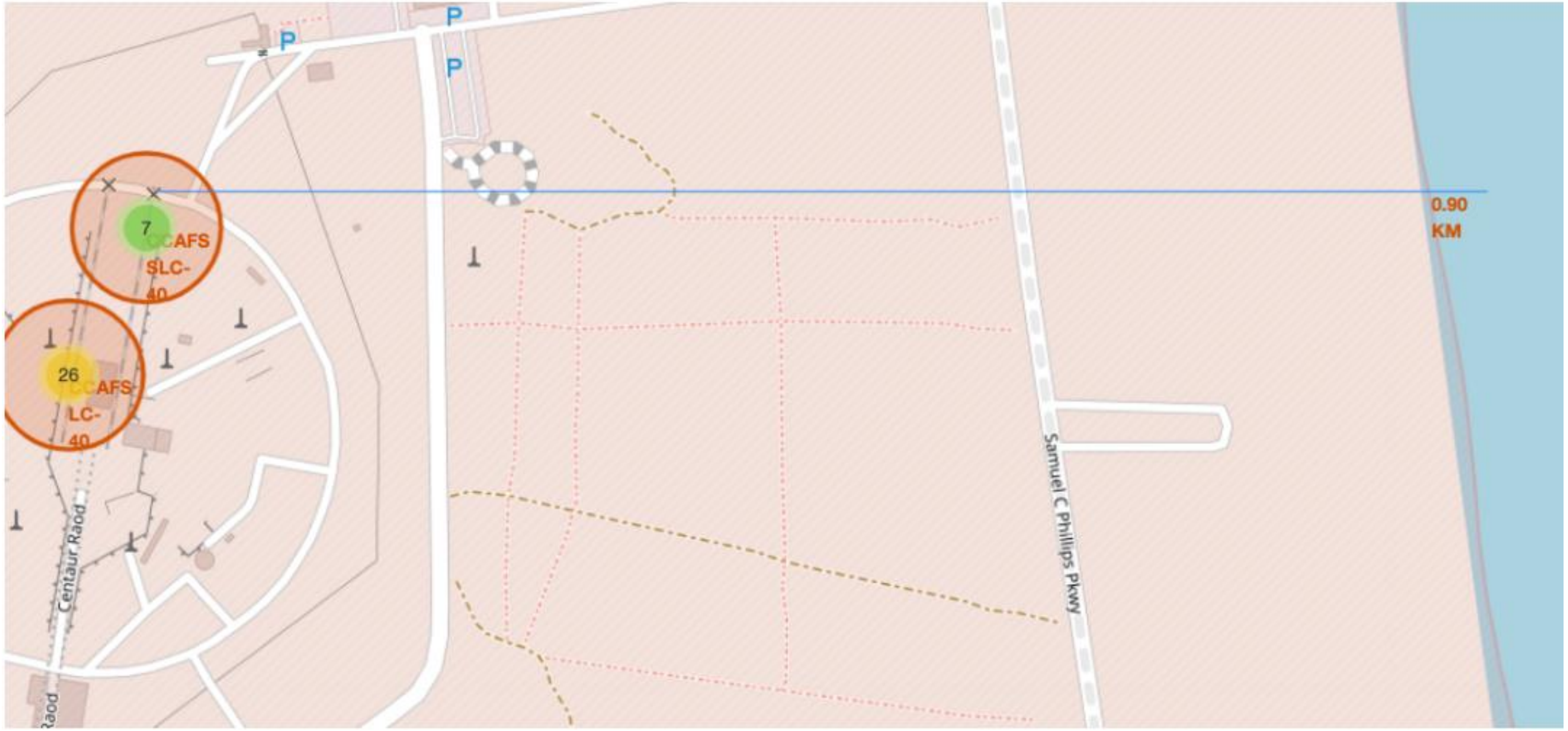
RESULTS: MAP WITH FOLIUM

- The figure below visualizes the launch sites and number of successful (green) and failed (red) launches:



RESULTS: MAP WITH FOLIUM

- The launch sites are in very close proximity to the coast. The figure below depicts the distance between CCAFS SLC 40 and the coastline:

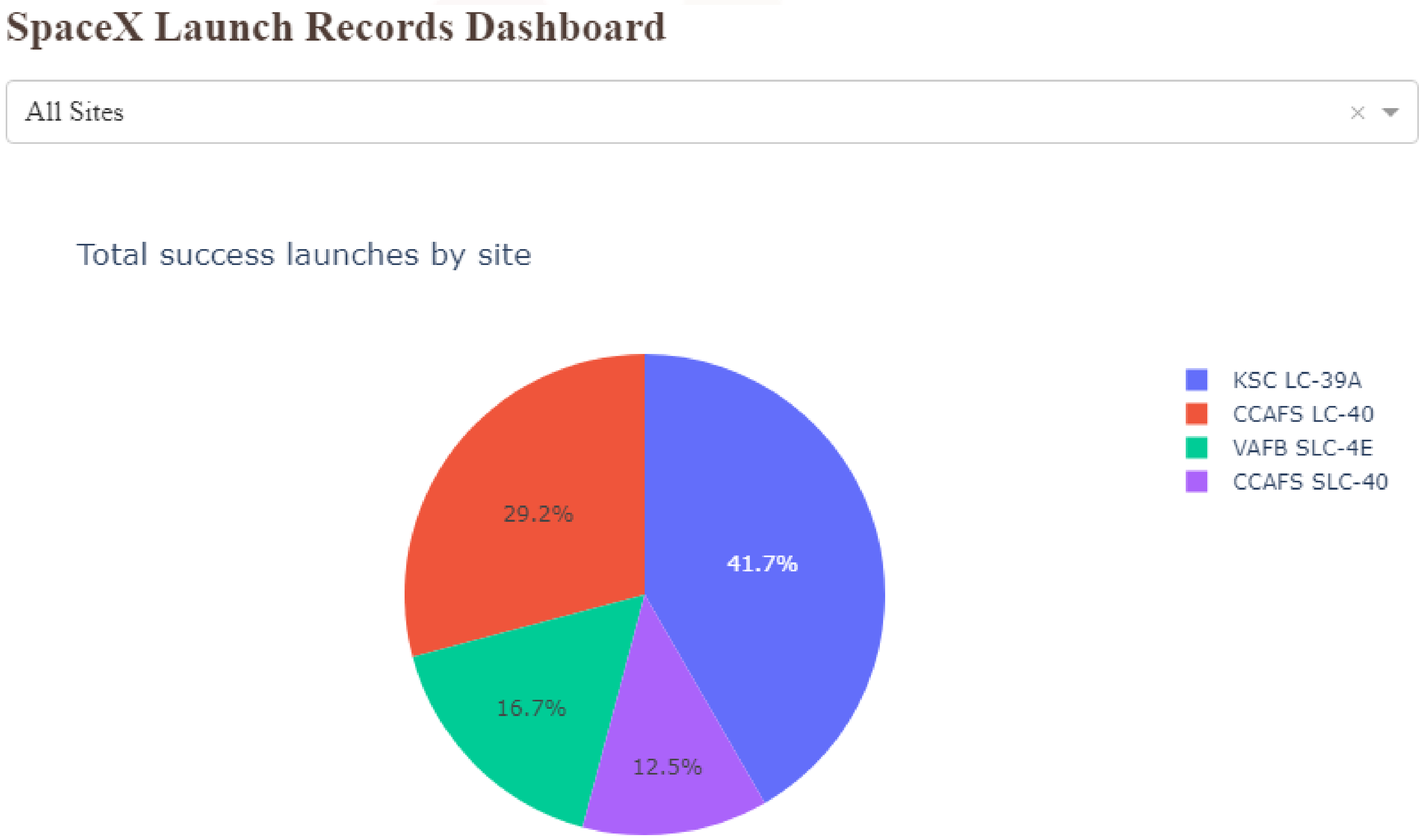


DASHBOARD

GitHub link of the Cognos dashboard:

https://github.com/mathias-wambeke/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

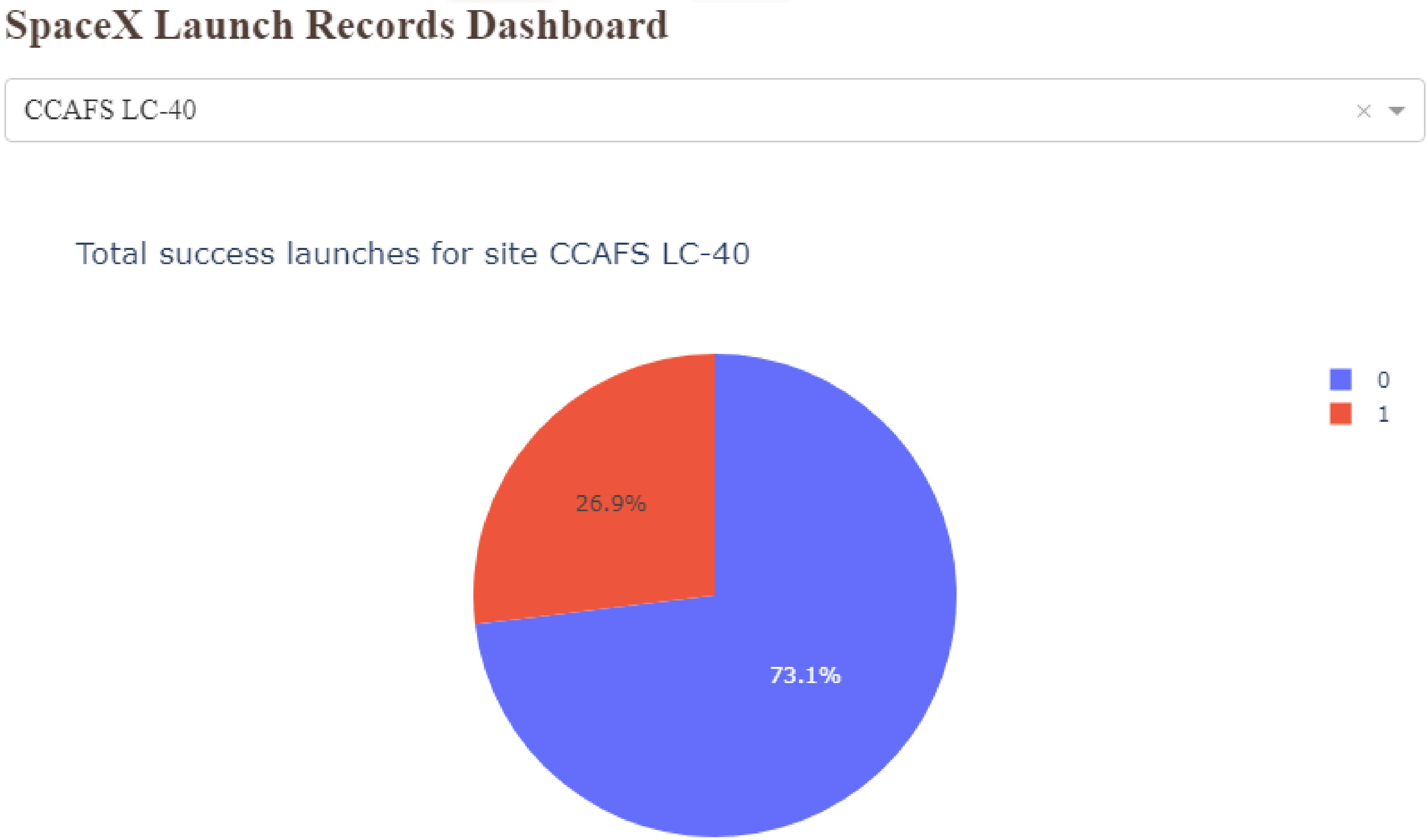
DASHBOARD TAB 1



SpaceX Launch Records Dashboard

All Sites

DASHBOARD TAB 2

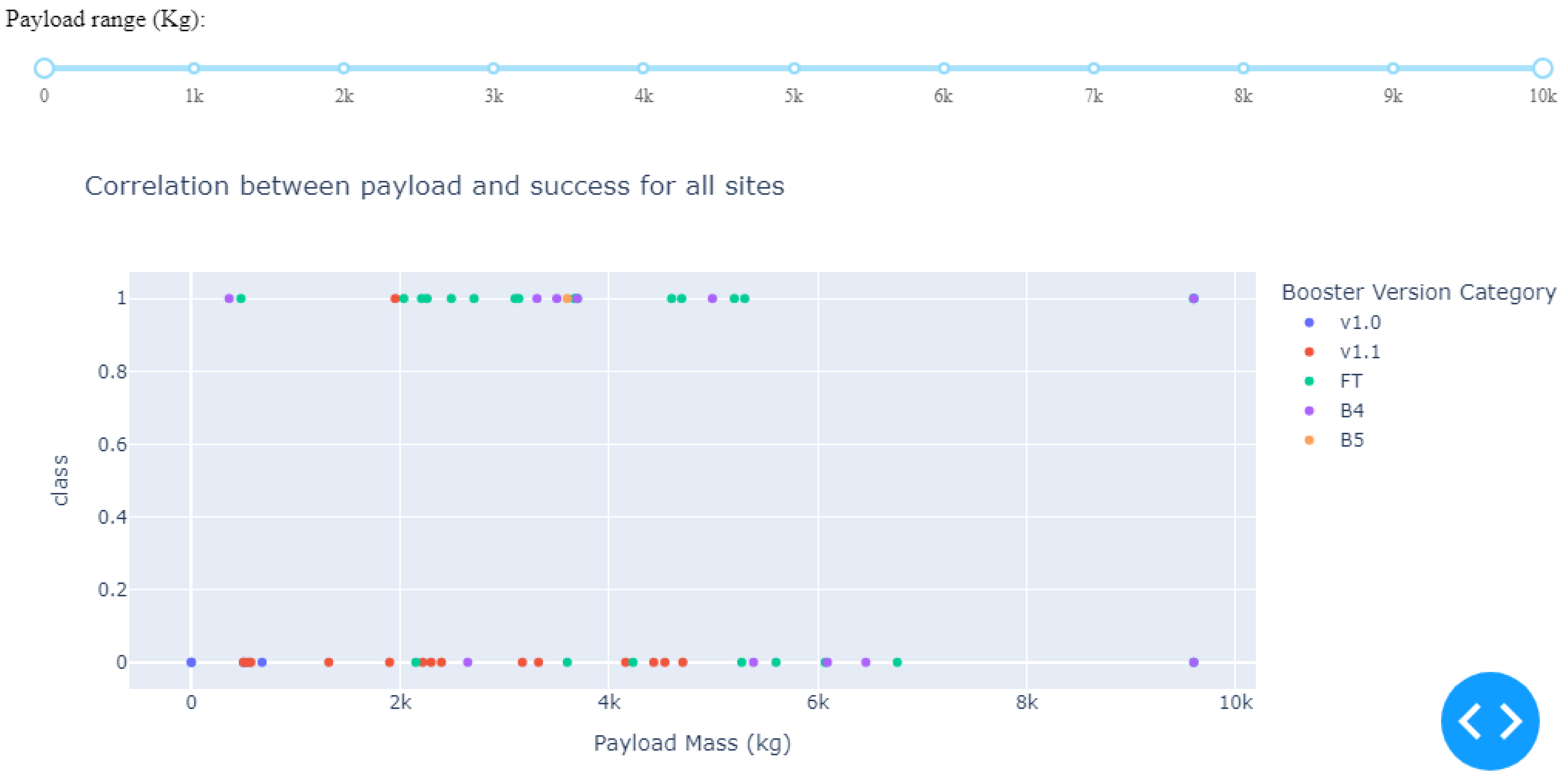


SpaceX Launch Records Dashboard

CCAFS LC-40

Total success launches for site CCAFS LC-40

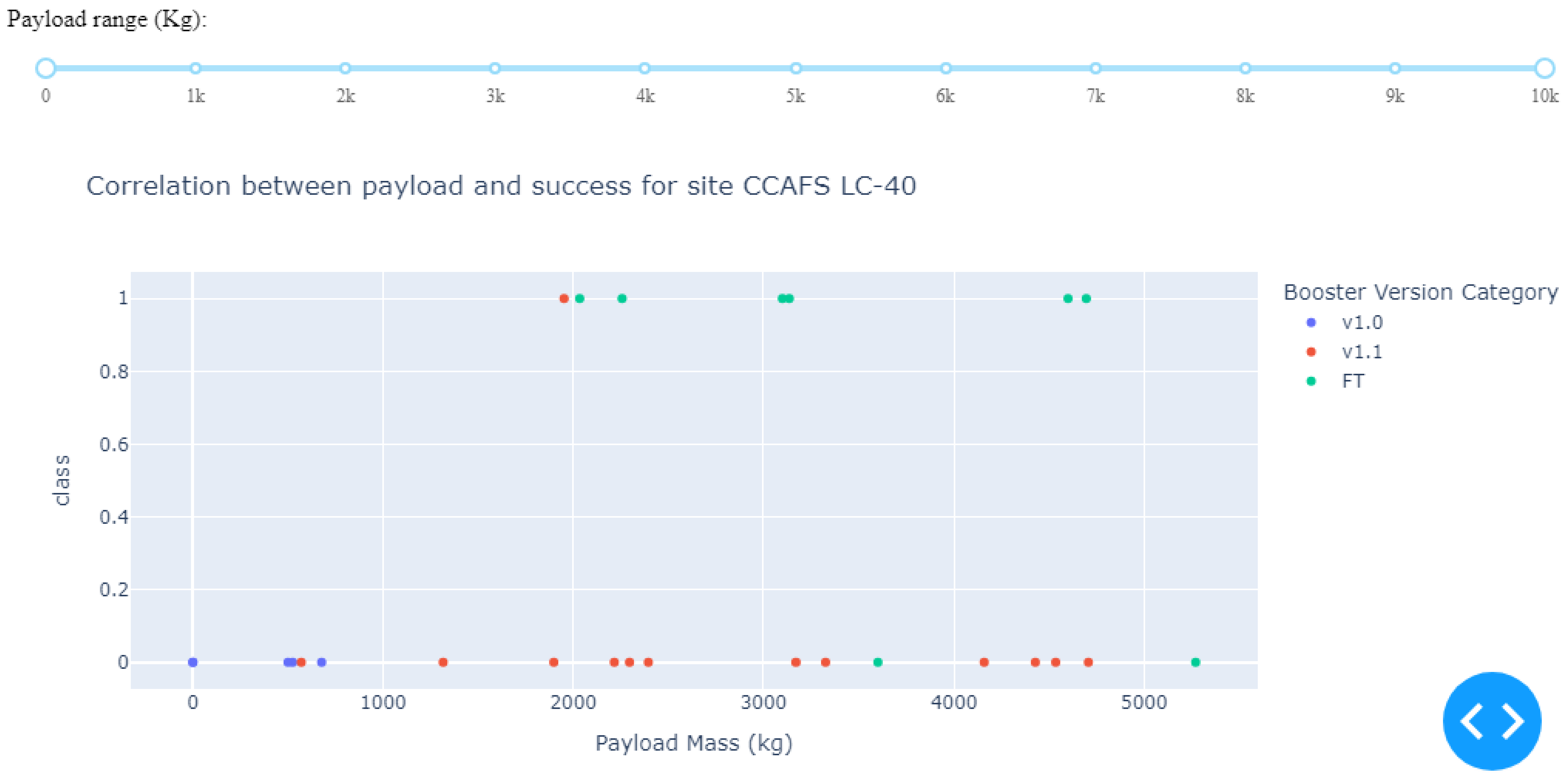
DASHBOARD TAB 3



Payload range (Kg):

Correlation between payload and success for all sites

DASHBOARD TAB 4



Payload range (Kg):

Correlation between payload and success for site CCAFS LC-40

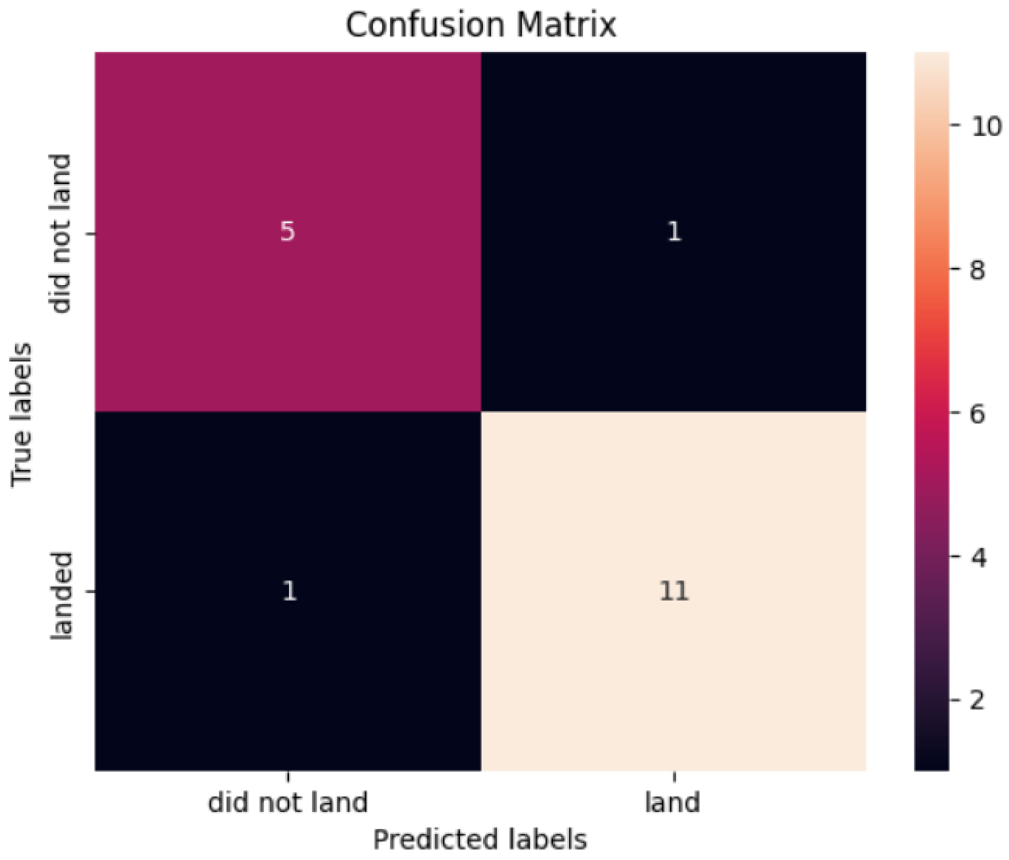
RESULTS: PREDICTIVE ANALYSIS

- The table below presents the accuracy score for the classification models
- The classification tree model presents the highest accuracy for both the train & test set

Model	Accuracy train	Accuracy test
Logistic Regression	84.6%	83.3%
SVM	84.8%	83.3%
Classification Tree	87.5%	88.8%
K Nearest Neighbors	84.8%	83.3%

RESULTS: PREDICTIVE ANALYSIS

- The confusion matrix for the classification tree, the best performing model, is presented below:



DISCUSSION

FINDINGS, IMPLICATIONS & INSIGHTS

- We observe some clear time effects:
  - As the flight number increases, the first stage is more likely to land successfully
  - The success rate since 2013 kept increasing till 2020
- A higher payload, at a particular point in time, appears to be related to a lower success rate
  - The overall relationship between payload & success rate is however not very clear and may be confounded by the observation that, over time, higher payloads are used
- The GTO orbit has the lowest success rate
  - We also find that the relationship between flight number & success rate is less clear when in GTO orbit

FINDINGS, IMPLICATIONS & INSIGHTS

- Different launch sites have different success rates:
  - KSC LC-39A has the highest success rate at 77%
  - CCAFS LC-40, which was predominantly used during the first years, has the lowest success rate at 27%
- Different booster versions have different success rates:
  - Booster versions V.1 & V1.1., used during the first years, have the lowest success rate at 0% & 7% respectively
  - Booster versions B4 & FT, used during later years, have the highest success rate at 55% & 67% respectively
- The main implication & insight appears that, over time, the success rate increases very significantly:
  - While the success rate is 0% up to 2013, it reaches >80% as of 2019
  - A higher success rate is thus mainly achieved through experience and is likely related to continuing investments in e.g. booster technology.

CONCLUSION

- This research aimed at predicting successful landings of the Falcon 9 first stage
- We observe clear time effects: as the flight number increases, the first stage is more likely to land successfully. This time effect also explains the higher success rates of launch sites and booster versions that were used during later years
- At a particular point in time, a higher payload appears to be related to a lower success rate
- The GTO orbit has the lowest success rate
- Overall, we conclude that a higher success rate is mainly achieved through experience over time and continuing improvements in technology

APPENDIX

Relevant links & sources:

- SpaceX API:
  - https://api.spacexdata.com/v4/launches/past
- List of Falcon 9 and Falcon Heavy launches:
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Capstone project:
  - GitHub - mathias-wambeke/Applied-Data-Science-Capstone: Applied Data Science Capstone
- Cognos dashboard:
  - https://github.com/mathias-wambeke/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py