IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emad Eldeen Elsayed Mohammed
10/06/2023

Outline

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

Executive Summary

- In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.
- The main steps in this project include:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- SpaceY is a recently established player in the commercial rocket launch industry, aiming to compete with SpaceX by bidding against them for launch contracts.
- SpaceX advertises launch services starting at $67 million, which includes reserving fuel for reusing the first-stage rocket booster.
- According to public statements from SpaceX, the cost of building a first-stage Falcon 9 rocket booster is estimated to be over $15 million, not including R&D expenses or profit margin.
- The report shows that by considering mission parameters like payload mass and desired orbit, the models were able to predict the first-stage rocket booster landing with an accuracy level of 83.3%.
- Using first-stage landing predictions as a cost proxy for launches, SpaceYcan make more informed bids against SpaceX.

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully.
  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.
- Most unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean.
- The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which includes its payload mass, orbit type, launch site, and so on, will the first stage of the rocket land successfully?

Section 1

Methodology

METHODOLOGY

- The overall methodology includes:
  1. Data collection, wrangling, and formatting, using:
     - SpaceX API
     - Web scraping
  2. Exploratory data analysis (EDA), using:
     - Pandas and NumPy
     - SQL
  3. Data visualization, using:
     - Matplotlib and Seaborn
     - Folium
     - Dash
  4. Machine learning prediction, using
     - Logistic regression
     - Support vector machine (SVM)
     - Decision tree
     - K-nearest neighbors (KNN)

METHODOLOGY

Data collection and wrangling

- SpaceX API
  - The API used is https://api.spacexdata.com/v4/rockets/.
  - The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
  - Every missing value in the data is replaced the mean the column that the missing value belongs to.
  - We end up with 90 rows or instances and 17 columns or features. The picture below shows the first few rows of the data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

METHODOLOGY

Data collection and wrangling

- Web scraping
  - The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - The website contains only the data about Falcon 9 launches.
  - We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

METHODOLOGY

Data collection and wrangling

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.

METHODOLOGY

Exploratory Data Analysis (EDA)

- Pandas and NumPy
  - Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:
    - The number of launches on each launch site
    - The number of occurrence of each orbit
    - The number and occurrence of each mission outcome
- SQL
  - The data is queried using SQL to answer several questions about the data such as:
    - The names of the unique launch sites in the space mission
    - The total payload mass carried by boosters launched by NASA (CRS)
    - The average payload mass carried by booster version F9 v1.1

METHODOLOGY

Data Visualization

- Matplotlib and Seaborn
  - Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.
  - The plots and charts are used to understand more about the relationships between several features, such as:
    - The relationship between flight number and launch site
    - The relationship between payload mass and launch site
    - The relationship between success rate and orbit type
    - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/05-%20EDA%20Visualization.ipynb

- Folium
  - Functions from the Folium libraries are used to visualize the data through interactive maps.
  - The Folium library is used to:
    - Mark all launch sites on a map
    - Mark the succeeded launches and failed launches for each site on the map
    - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway
    - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/06-%206Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

METHODOLOGY

Data Visualization

- Dash
  - Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
  - Using a pie chart and a scatterplot, the interactive site shows:
    - The total success launches from each launch site
    - The correlation between payload mass and mission outcome (success or failure) for each launch site
    - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/07-%20spacex_dash_app.py

METHODOLOGY

Machine Learning Prediction

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase include the following steps:
  - Standardizing the data
  - Splitting the data into training and test data
  - Creating machine learning models, which include:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
  - Fit the models on the training set
  - Find the best combination of hyperparameters for each model
  - Evaluate the models based on their accuracy scores and confusion matrix
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/08-%20Machine%20Learning%20Prediction.ipynb

RESULTS

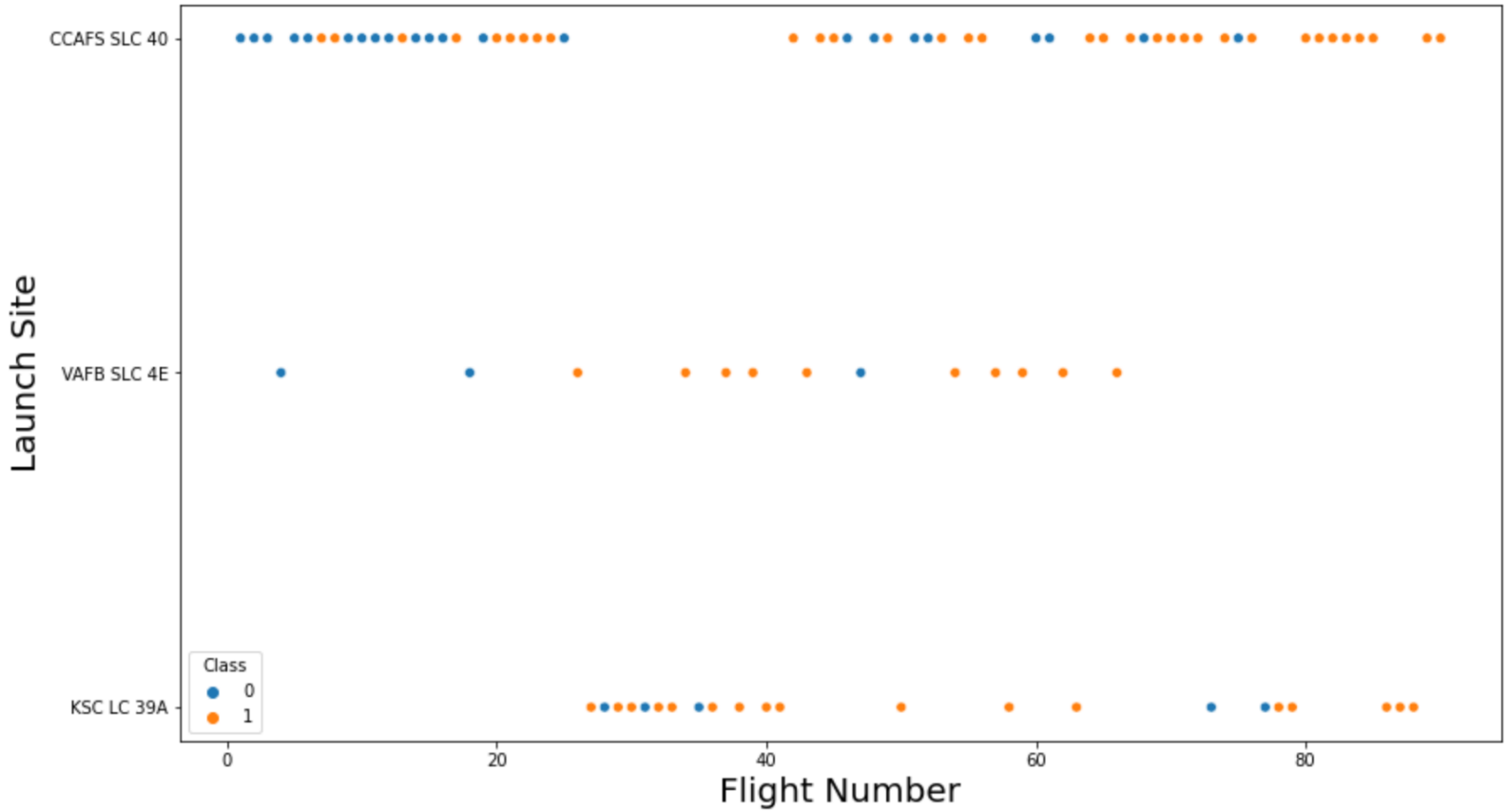
- The results are split into 5 sections:
  - SQL (EDA with SQL)
  - Matplotlib and Seaborn (EDA with Visualization)
  - Folium
  - Dash
  - Predictive Analysis
- In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

Section 2

Insights drawn from EDA

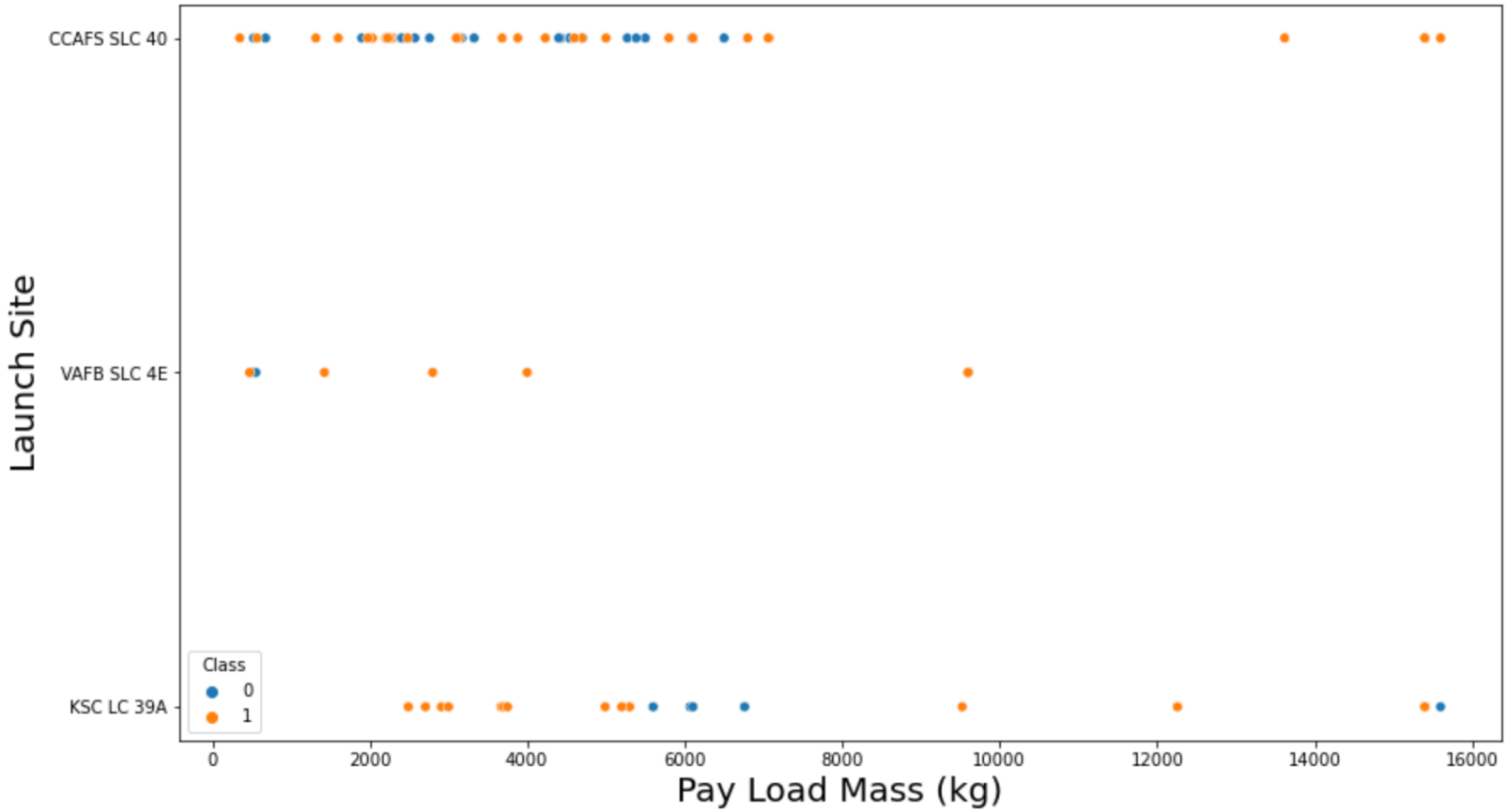
Flight Number vs. Launch Site

- The relationship between flight number and launch site



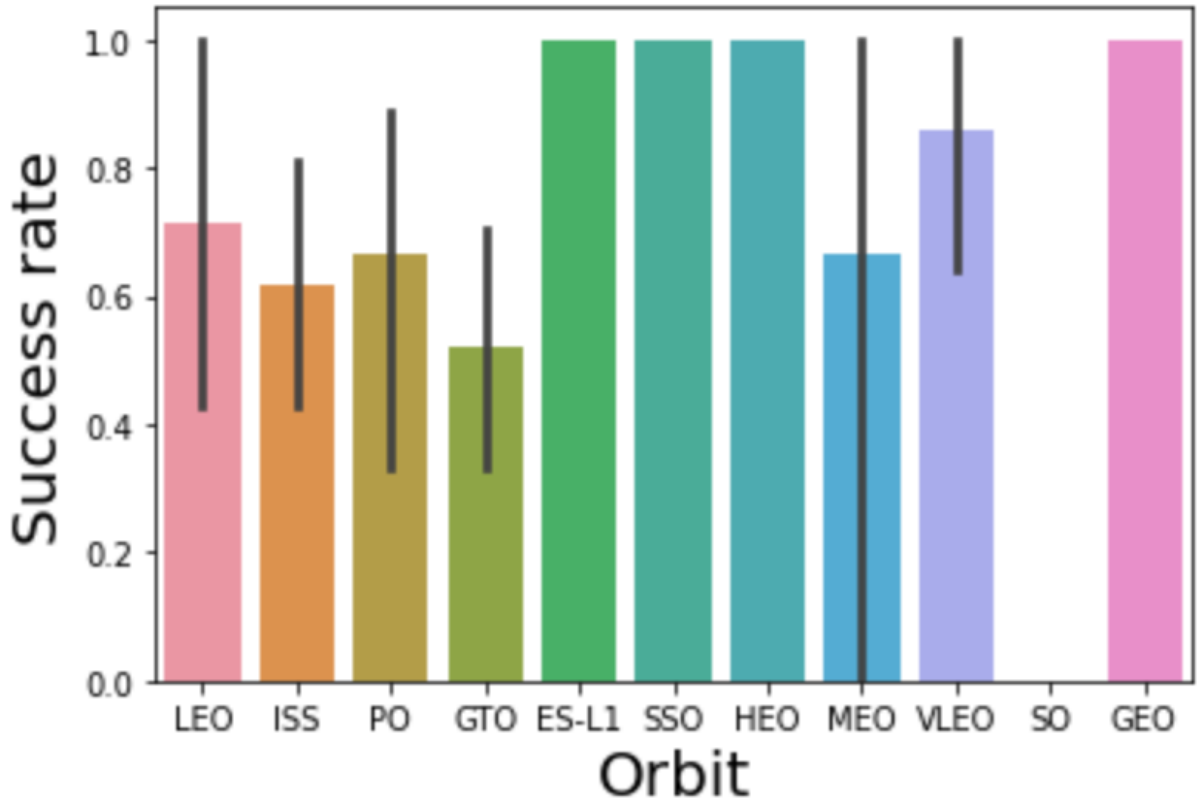
Payload vs. Launch Site

- The relationship between payload mass and launch site



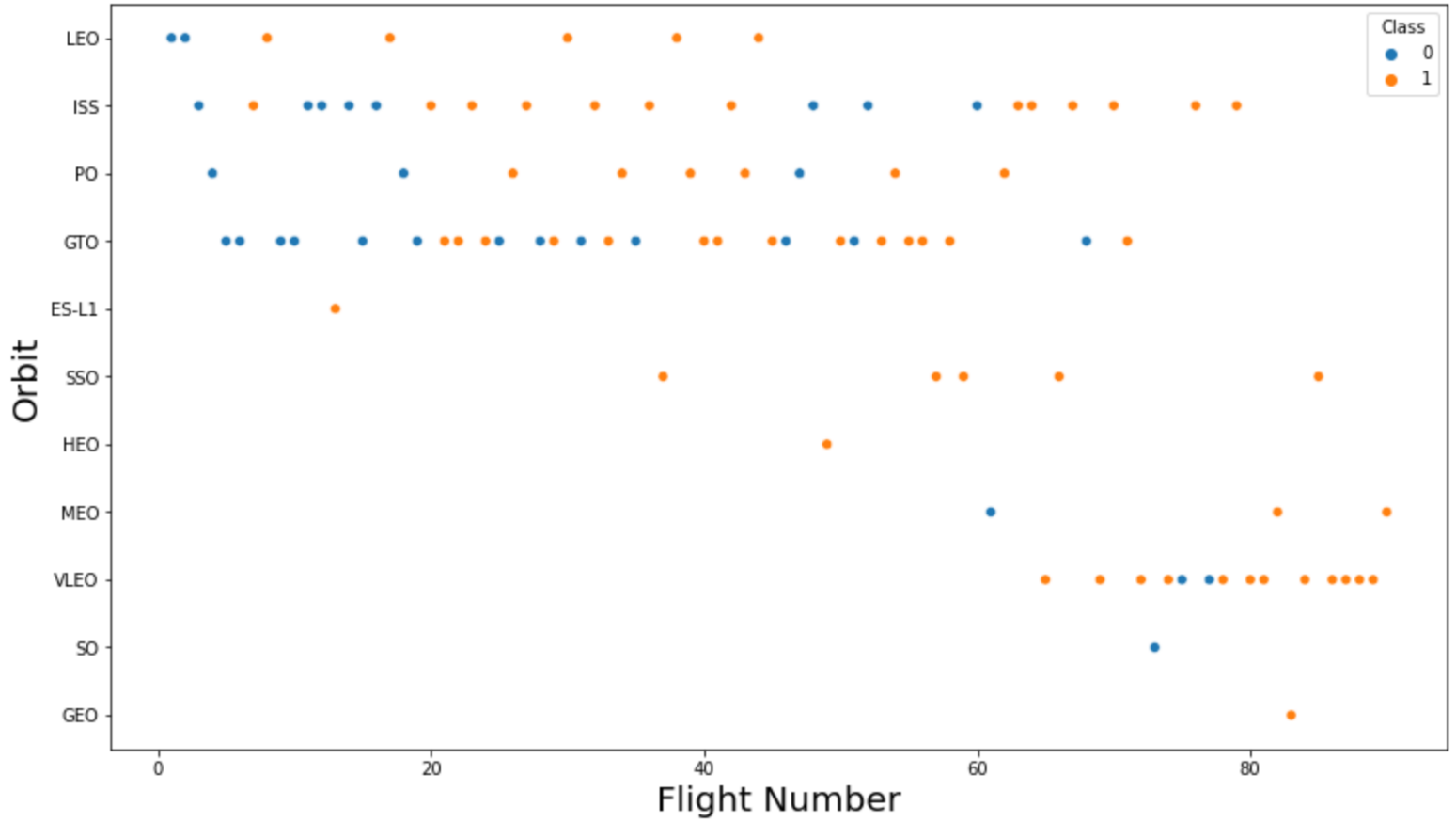
Success Rate vs. Orbit Type

- The relationship between success rate and orbit type



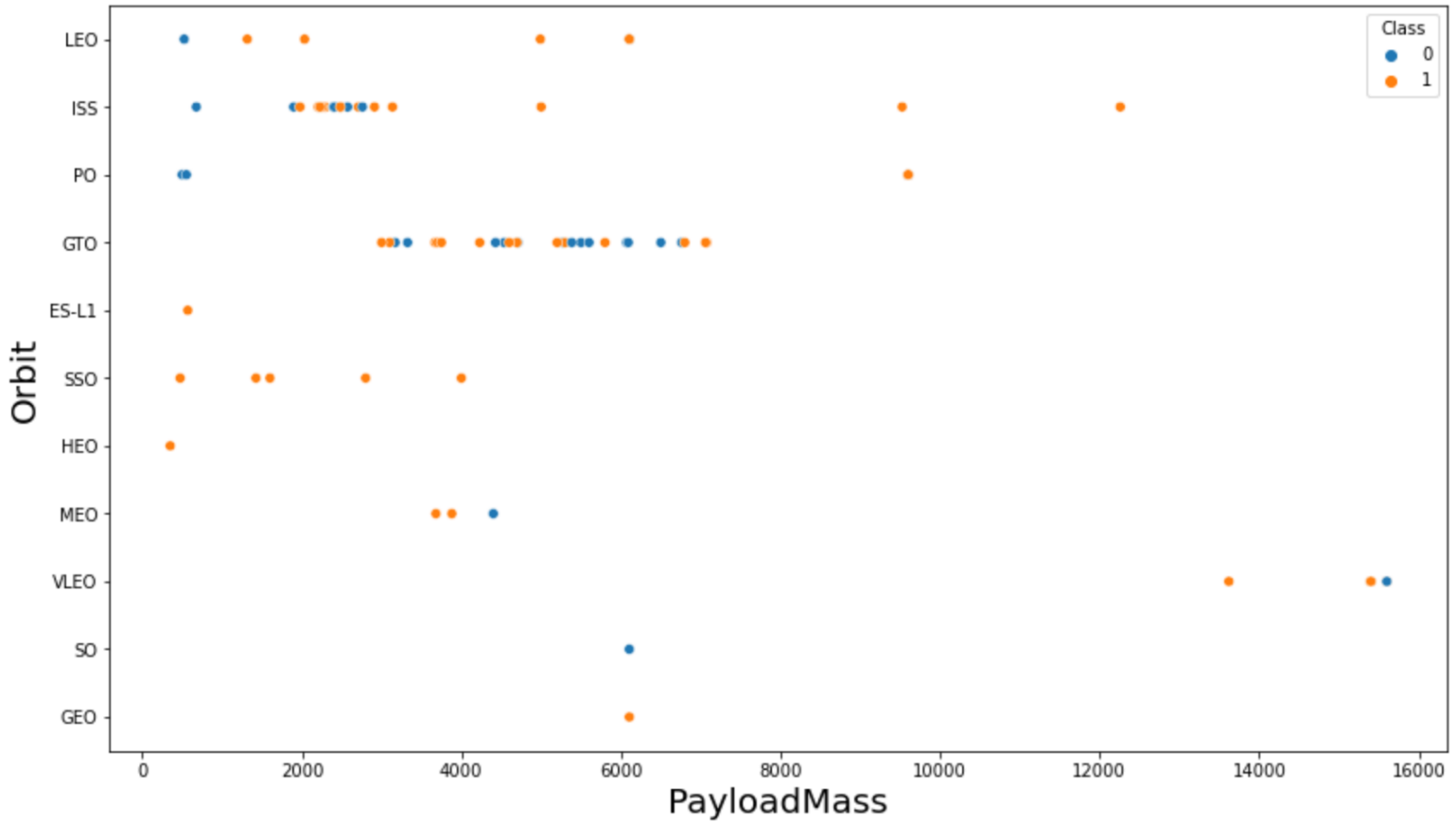
Flight Number vs. Orbit Type

- The relationship between flight number and orbit type



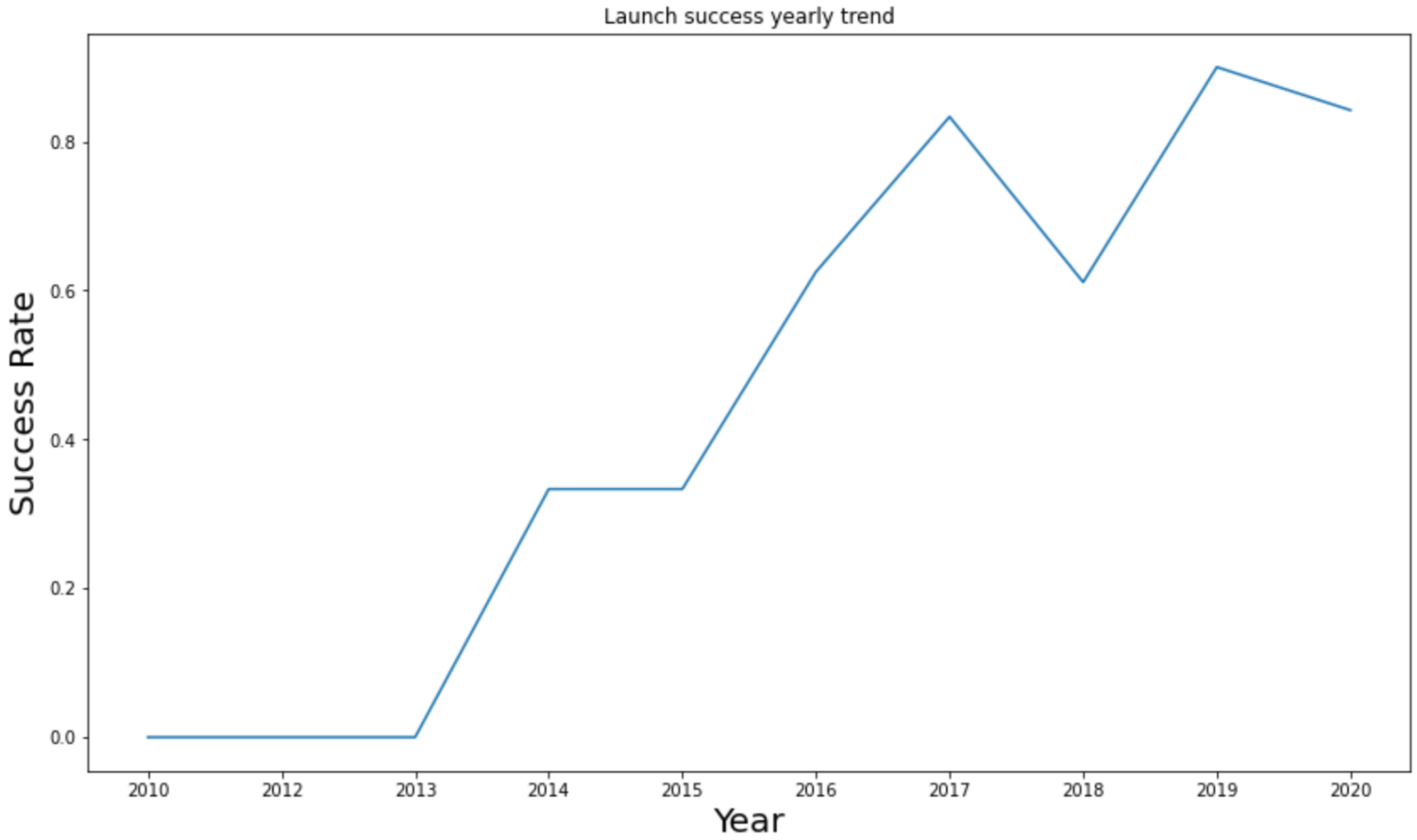
Payload vs. Orbit Type

- The relationship between payload mass and orbit type



Launch Success Yearly Trend

- The launch success yearly trend



All Launch Site Names

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

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with ‘CCA’
  - CCAFS LC-40
  - CCAFS SLC-40
  - Last launch from CCAFS LC-40 was 2016-08-14
  - First launch from CCAFS SLC-40 was 2017-12-15

Total Payload Mass

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

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1
    - 2534.67 Kilogram

First Successful Ground Landing Date

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

Date of first successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

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

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

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

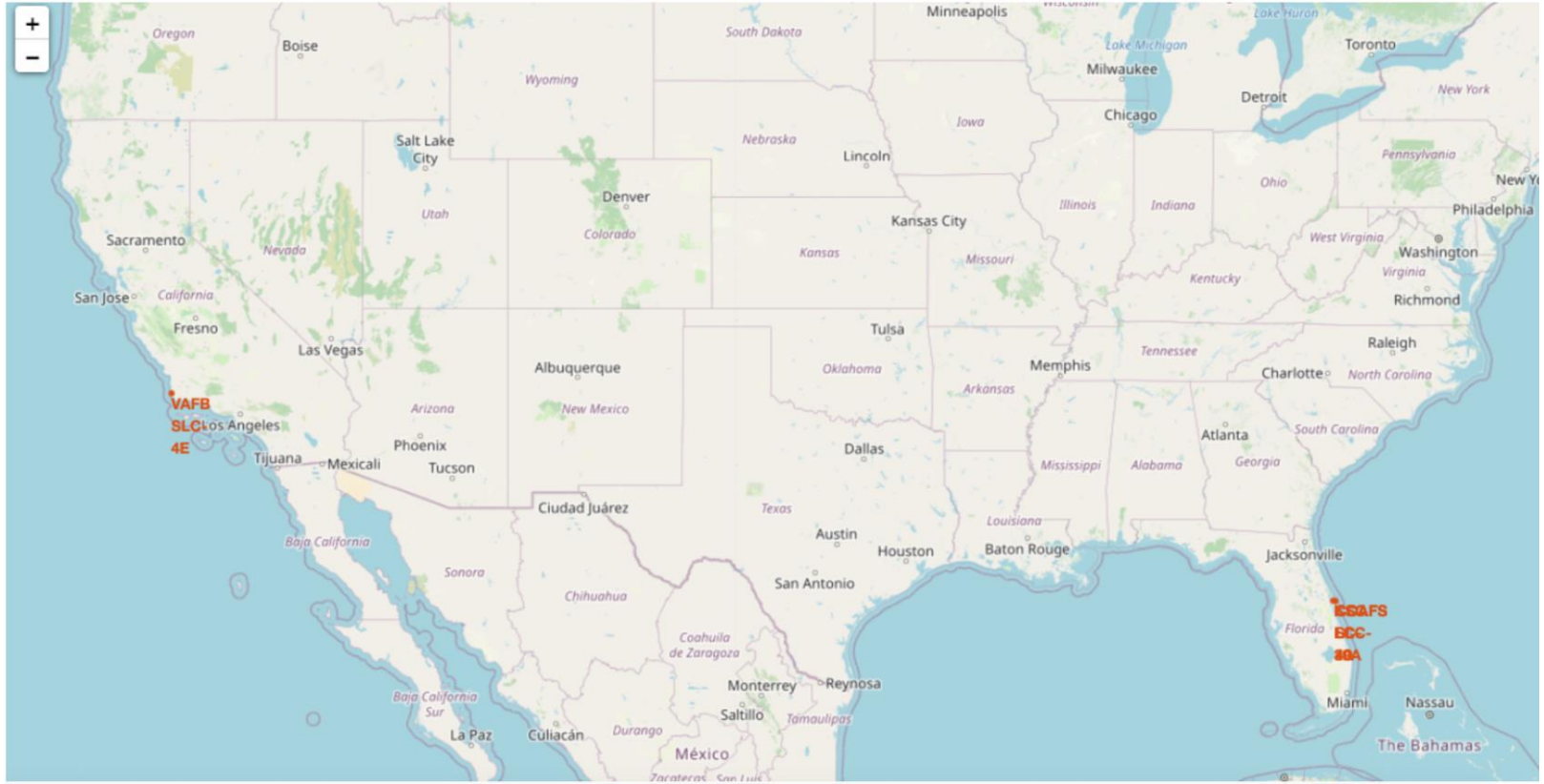
landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

RESULTS

Folium

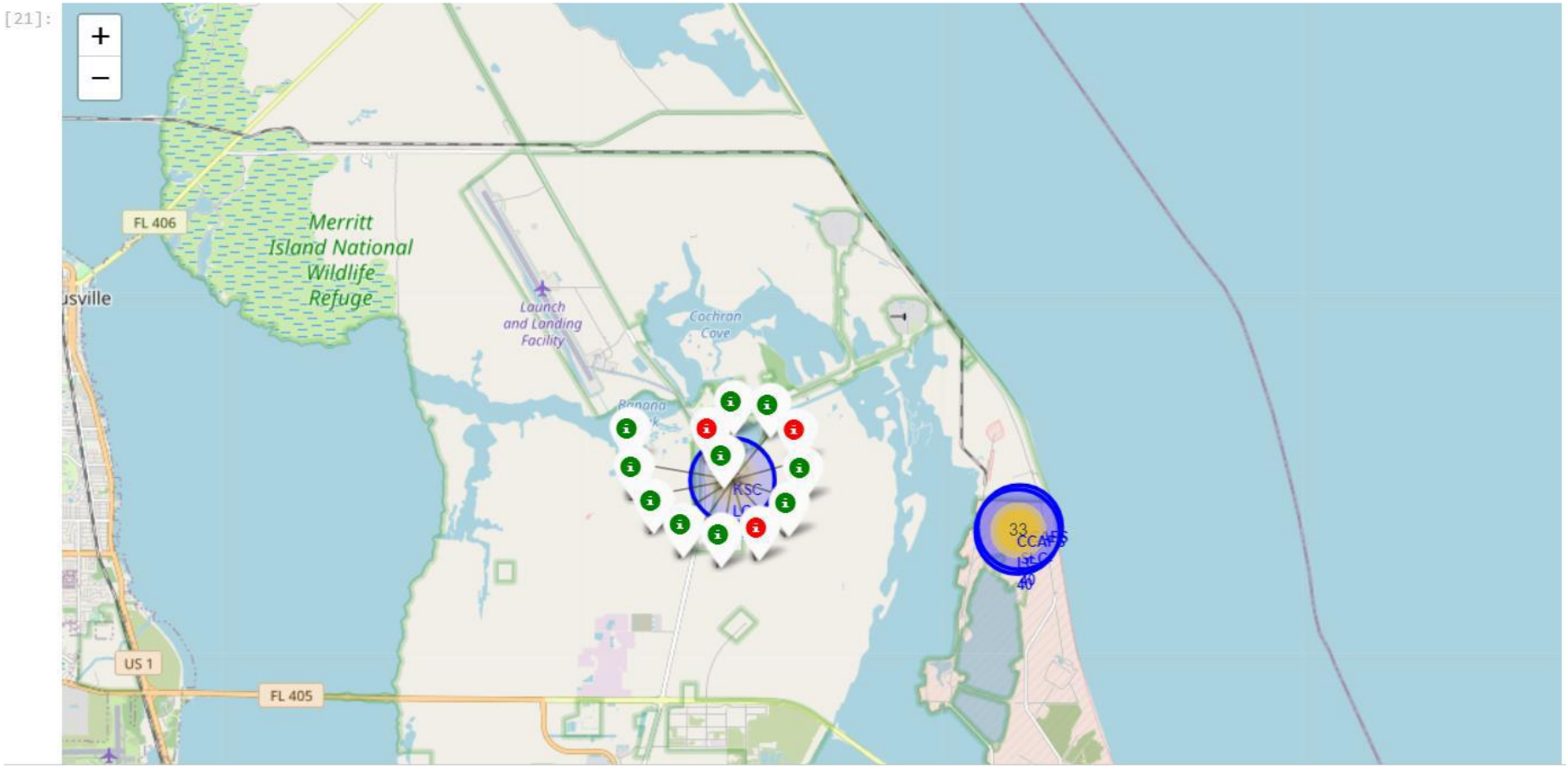
- All launch sites on map
- Visualizing the launch sites on a map highlights the importance of launch site proximity to the coast and equator using folium
-



RESULTS

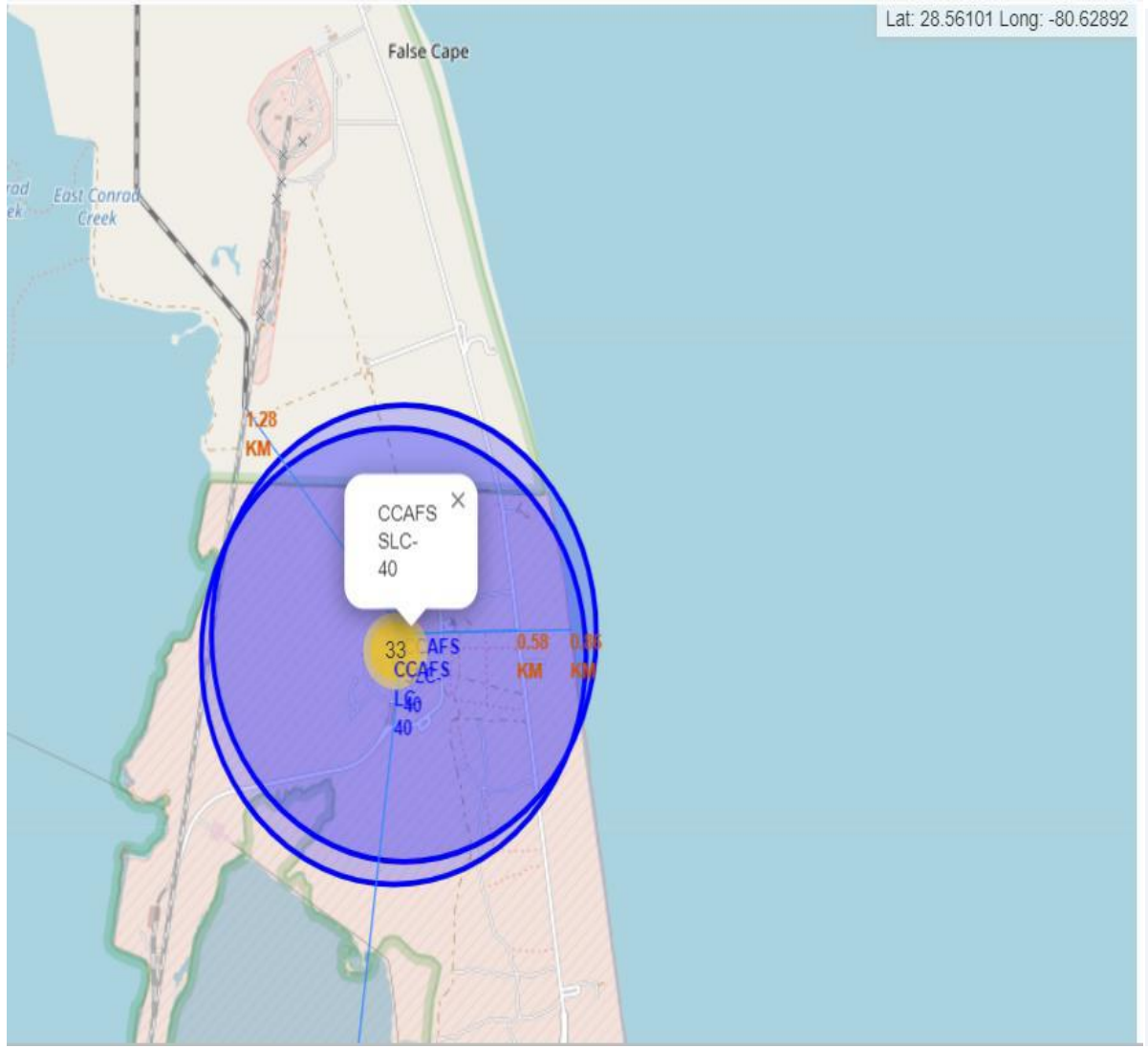
Folium

- Visualizing the booster landing outcomes for each launch site highlights which launch sites have relatively high success rates, namely KSC LC39A



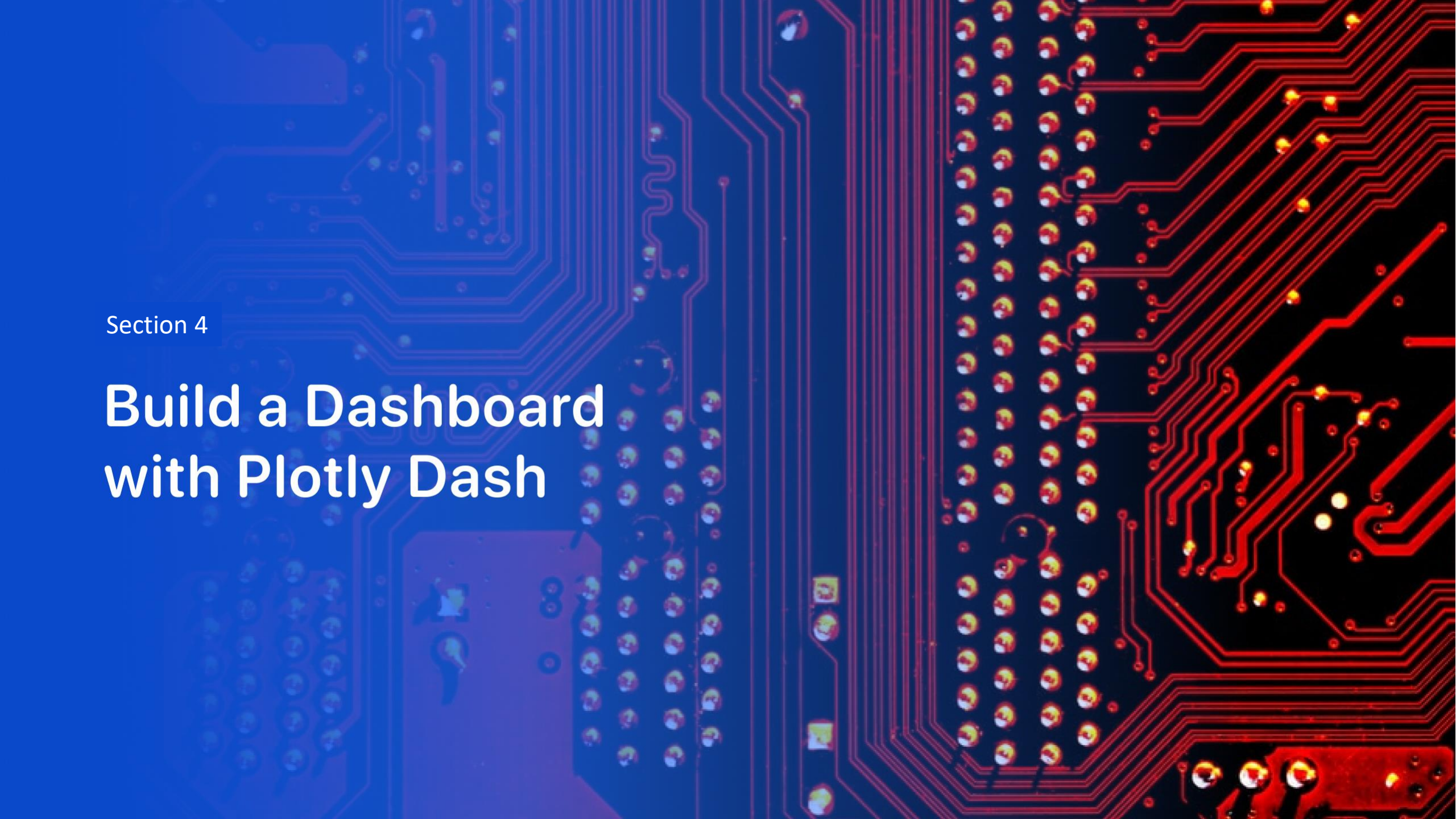
Folium

Visualizing the railway, highway, coastline, and city proximities for each launch site allows us to see how close each is present



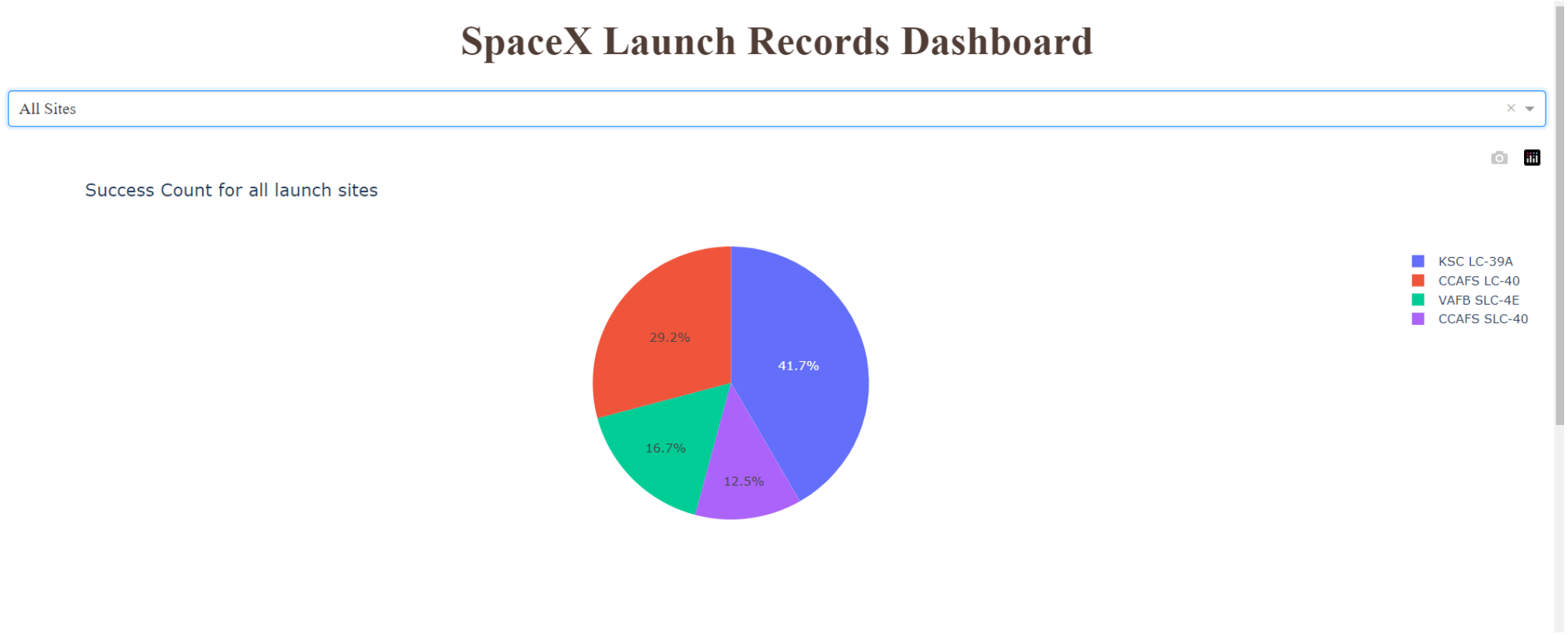
Proximities for CCAFS SLC-40:

- Railway: 1.28 km
  - transporting heavy cargo
- Highway: 0.58 km
  - transporting personal and equipment
- Coastline: 0.86 km
  - optionality to abort launch and attempt water landing
  - minimizing risk from falling debris
- City: 51.43 km
  - minimizing danger to population dense areas


Section 4
Build a Dashboard with Plotly Dash

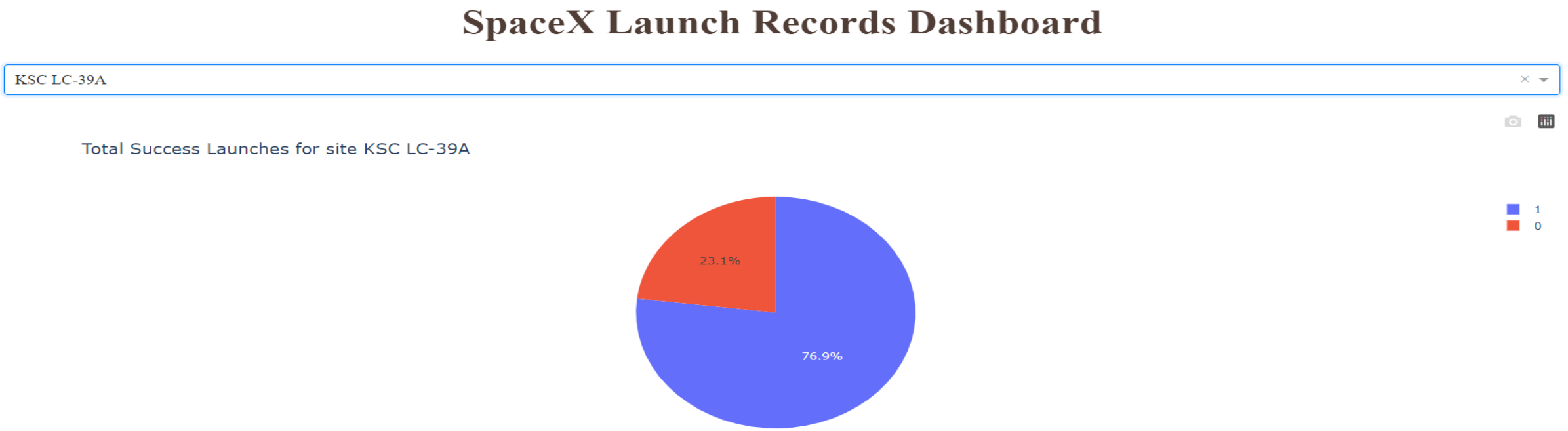
Launch Records Dashboard

- Drop-down menu to choose between all sites and individual launch sites.
- Color coded by launch sites.
- Pie chart showing booster landing success rate.



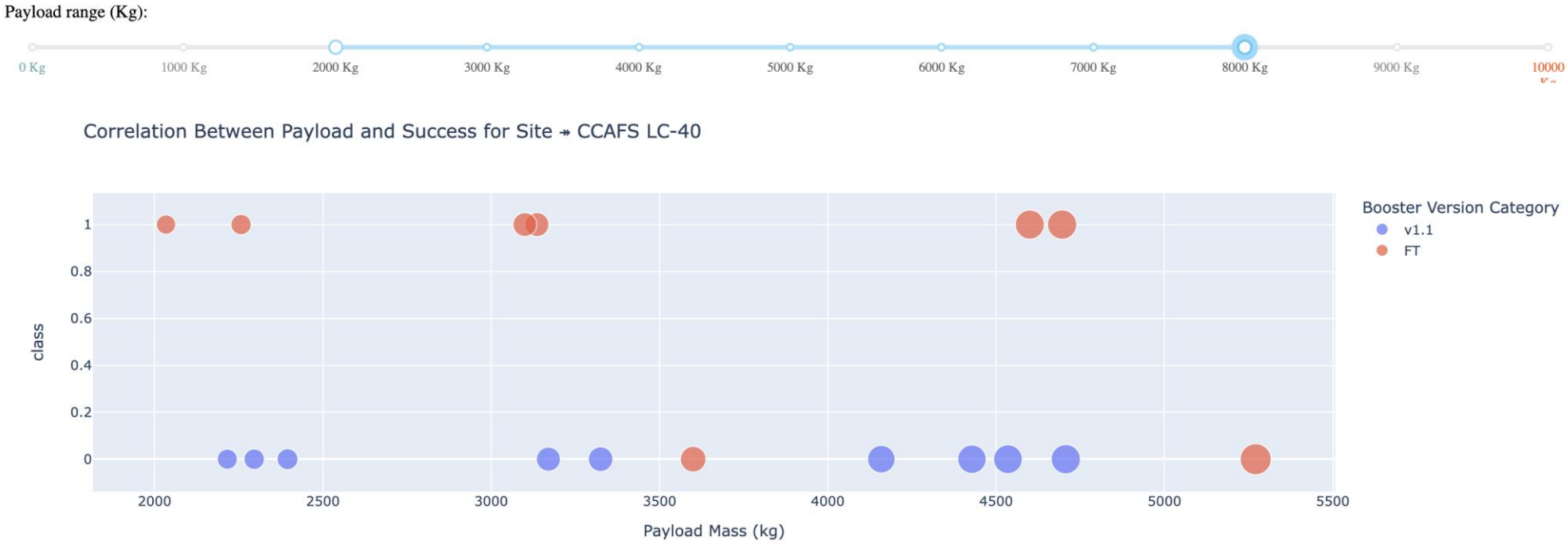
Dashboard Observations

- Enabling stakeholders to explore and manipulate the data in an interactive and real-time way
- KSC LC 39A has the highest booster landing success rate.



Dashboard Observations

- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.

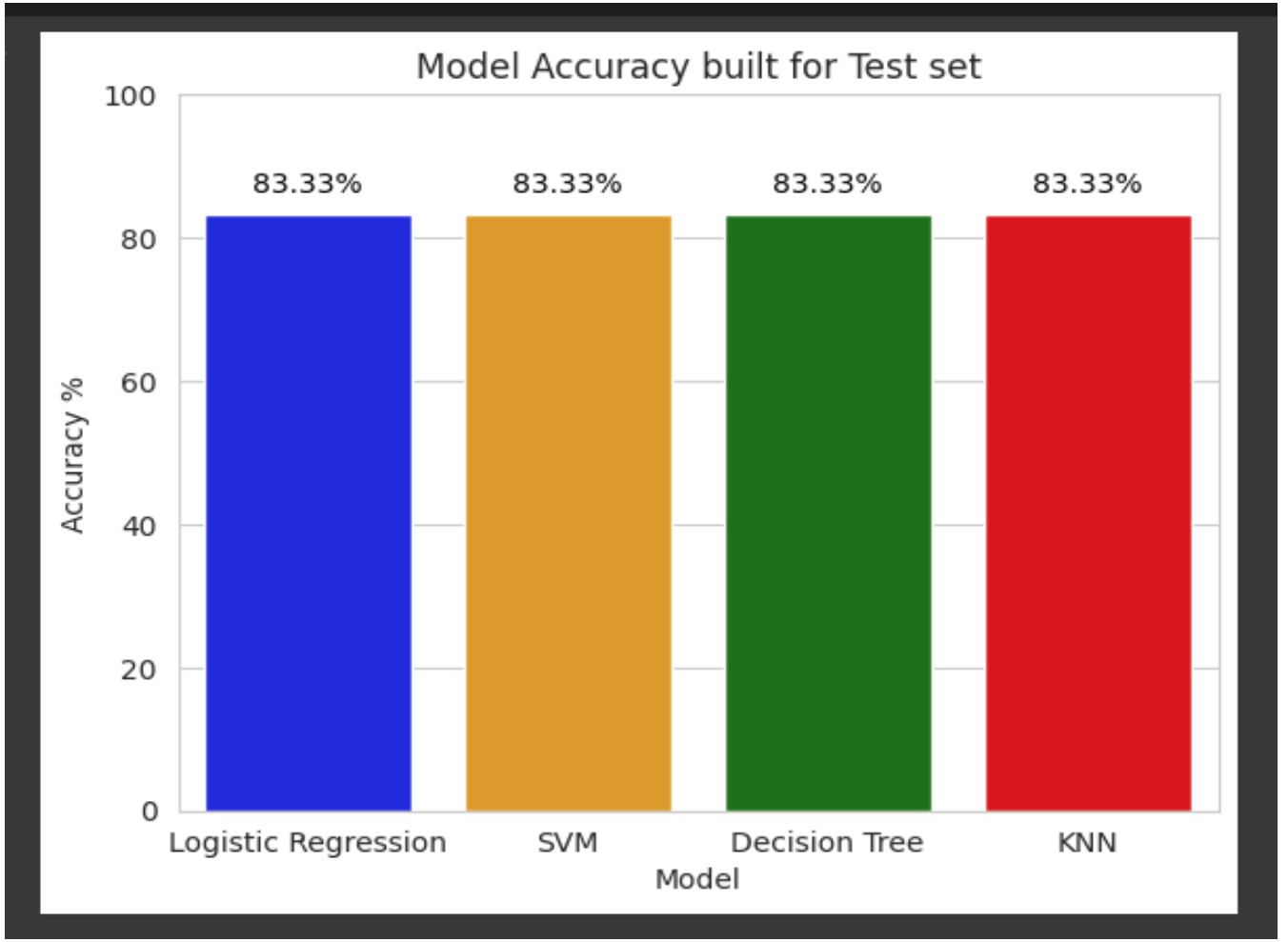


Section 5

Predictive Analysis (Classification)

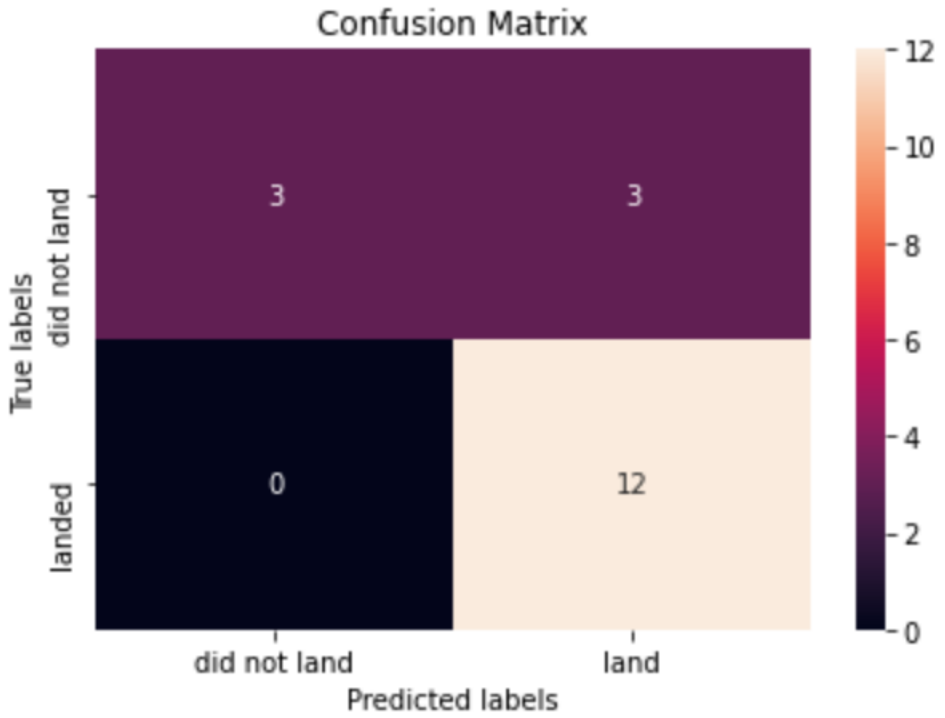
Classification Accuracy

- All the models built on the test set has an accuracy score of 83.33%.
- Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:
    1. Decision tree (GridSearchCV best score: 0.8892857142857142)
    2. K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
    3. Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
    4. Logistic regression (GridSearchCV best score: 0.8464285714285713)



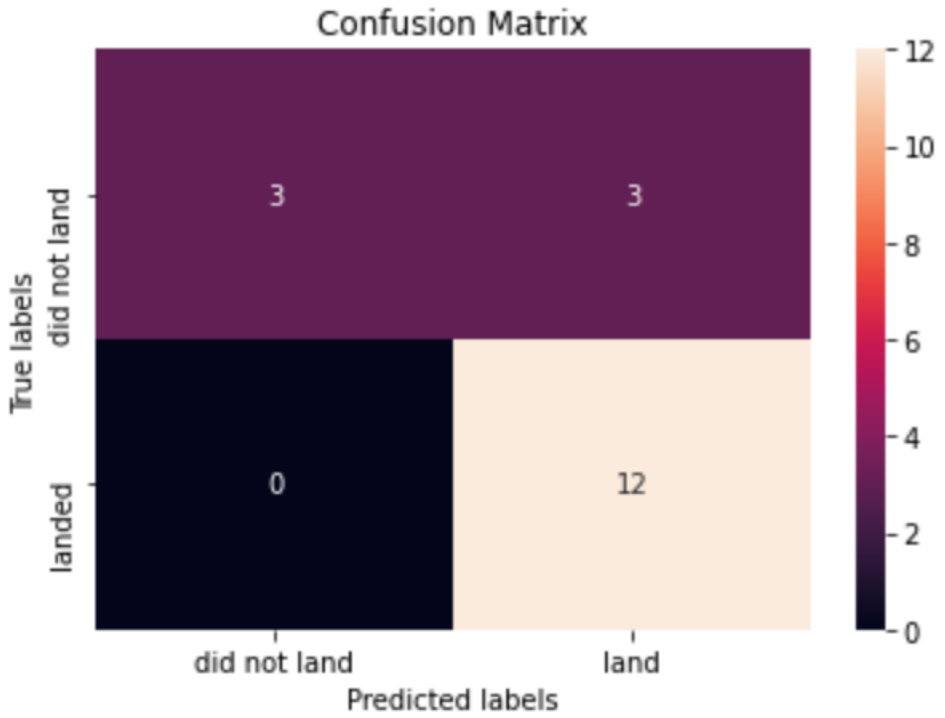
Confusion Matrix

- Logistic regression
  - GridSearchCV best score: 0.8464285714285713
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



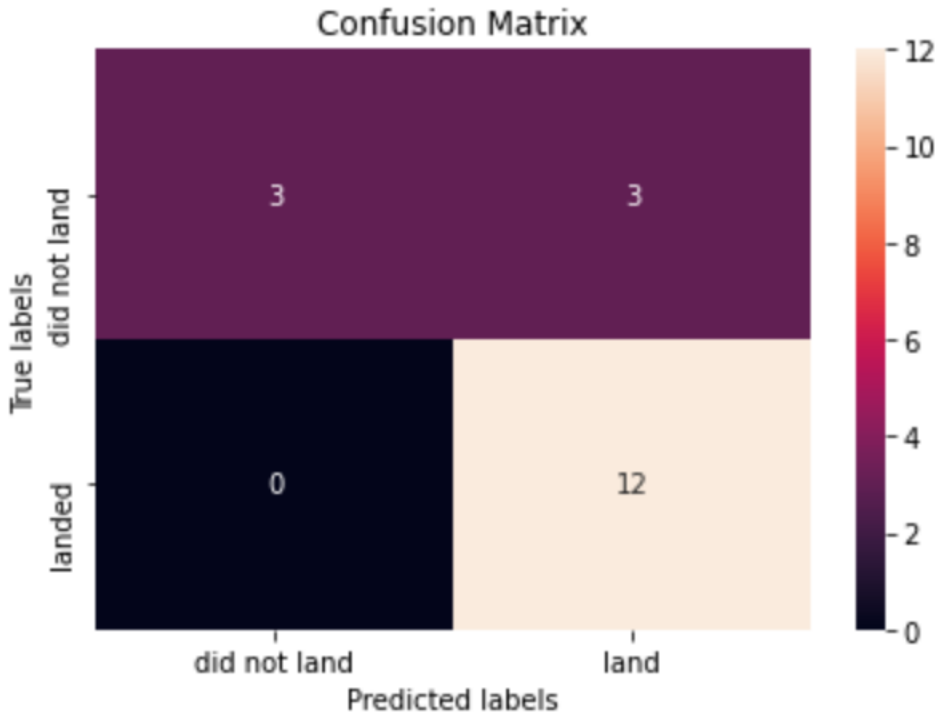
Confusion Matrix

- Support vector machine (SVM)
  - GridSearchCV best score: 0.8482142857142856
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



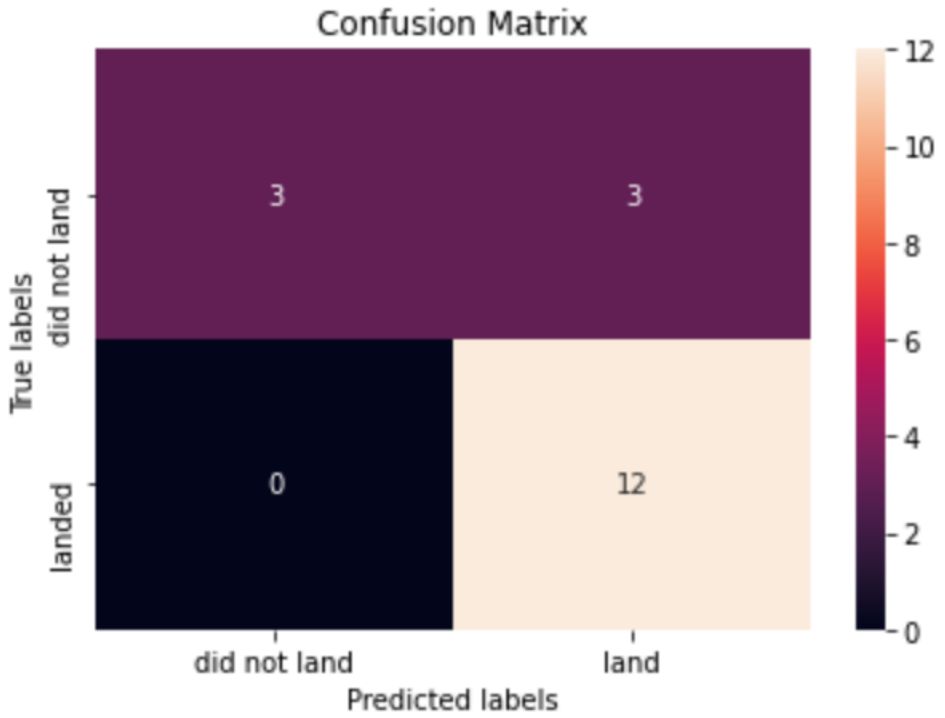
Confusion Matrix

- Decision tree
  - GridSearchCV best score: 0.8892857142857142
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



Confusion Matrix

- K nearest neighbors (KNN)
  - GridSearchCV best score: 0.8482142857142858
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



CONCLUSION

- This report provides models that SpaceY can use to make predictions about the successful landing of the 1st stage booster by SpaceX These models have an accuracy of 83 3 meaning that they can predict the outcome with a high level of confidence
- SpaceY will be able to make more informed bids against SpaceX This is because SpaceY will have a more accurate idea of when they can expect the SpaceX bid to include the cost of a sacrificed 1st stage booster, allowing them to better plan and allocate their resources
- In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch.
- Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.
- Several machine learning algorithms are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The predictive model produced by the decision tree algorithm performed the best among the 4 machine learning algorithms employed.

Appendix

- Python code snippets, SQL queries, charts, Notebook outputs, datasets analysis, and models:
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/01-%20Data%20Collection%20API.ipynb
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/02-%20Data%20Collection%20with%20Web%20Scraping.ipynb
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/03-%20EDA.ipynb
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/04-%20EDA%20with%20SQL.ipynb
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/05-%20EDA%20Visualization.ipynb
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/06-%206Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/07-%20spacex_dash_app.py
  - https://github.com/Emad-eldeen-Elsayed/IBM-Applied-Data-Science-Capstone-Project/blob/main/08-%20Machine%20Learning%20Prediction.ipynb

Thank you!