IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

smint
4 May 2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API, Web Scraping
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbours (KNN)

Data Collection

- API used: https://api.spacexdata.com/v4/rockets/.
  - This API provides data about types of rocket launches done by SpaceX, data is filtered to include only Falcon 9 launches
  - All missing values in the data is replaced with the mean
  - 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

Data Collection

- 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.
  - Remain with 121 rows or instances and 11 columns or features. Screenshot below shows first few rows of 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

EDA with Data Visualization

- Pandas and NumPy
  - Functions from Pandas and NumPy libraries were 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
  - Data was 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

Build an Interactive Map with Folium

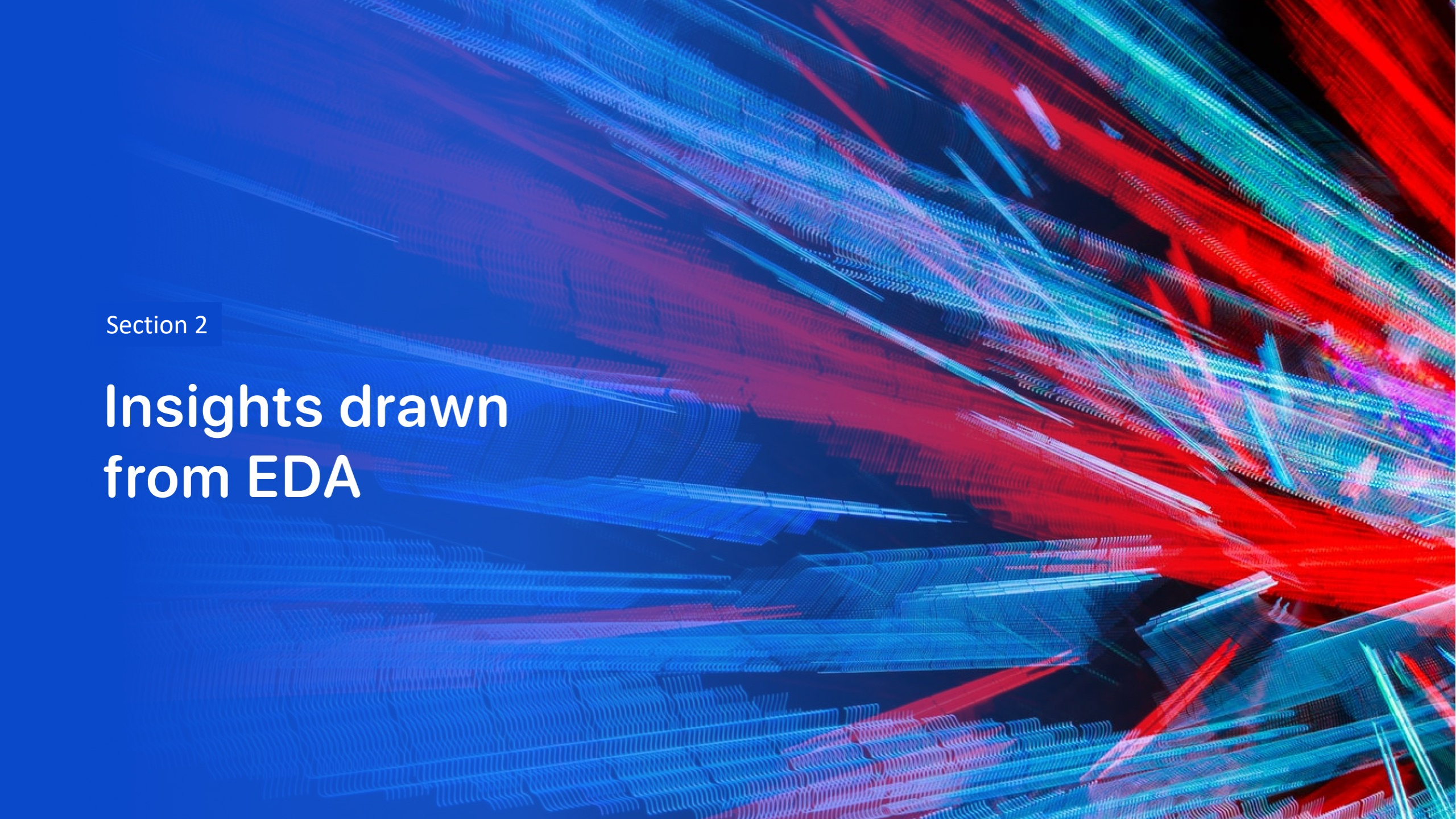
- Functions from the Folium libraries are used to visualize the data through interactive maps.
  - The Folium library was used to:
  - Mark all launch sites on a map
  - Mark the succeeded launches and failed launches for each site on the map
  - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

Predictive Analysis (Classification)

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase consist of 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

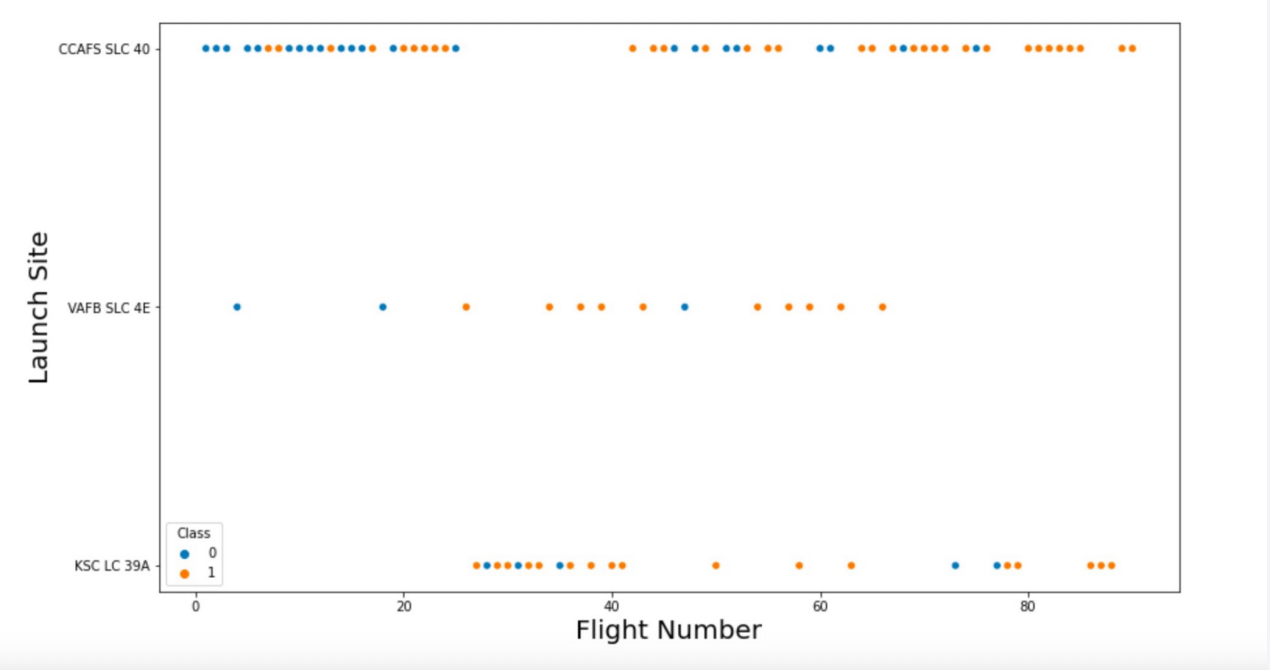
Results

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


Section 2
Insights drawn from EDA

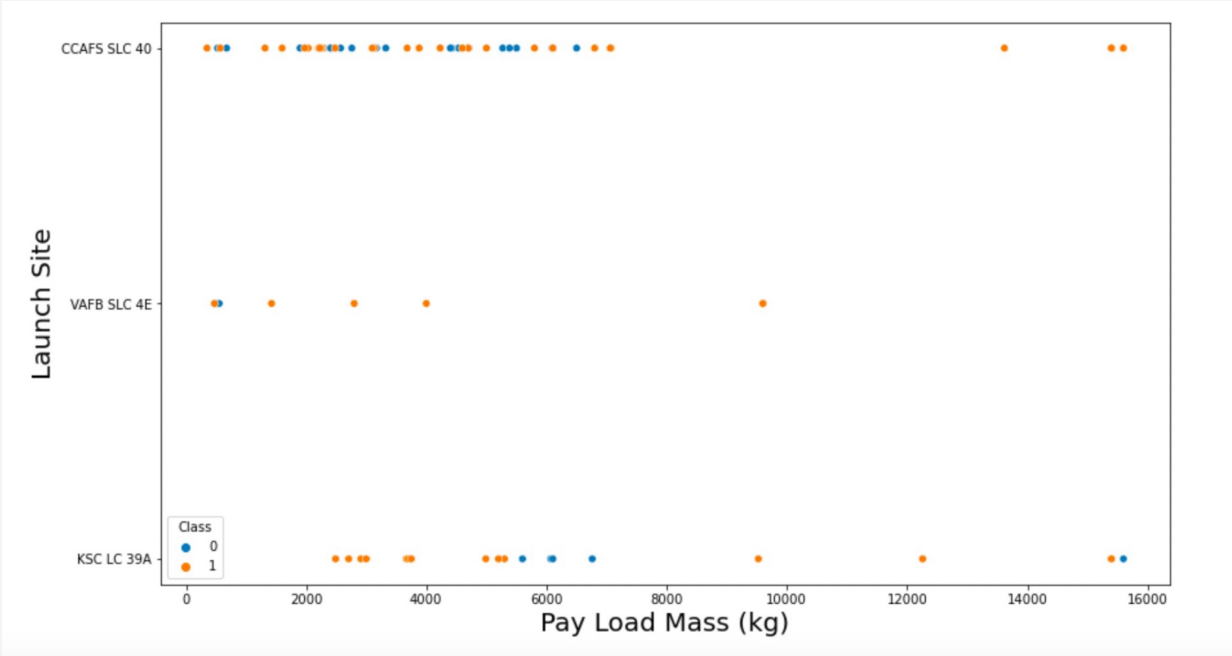
Flight Number vs. Launch Site

- Relationship between flight number and launch site



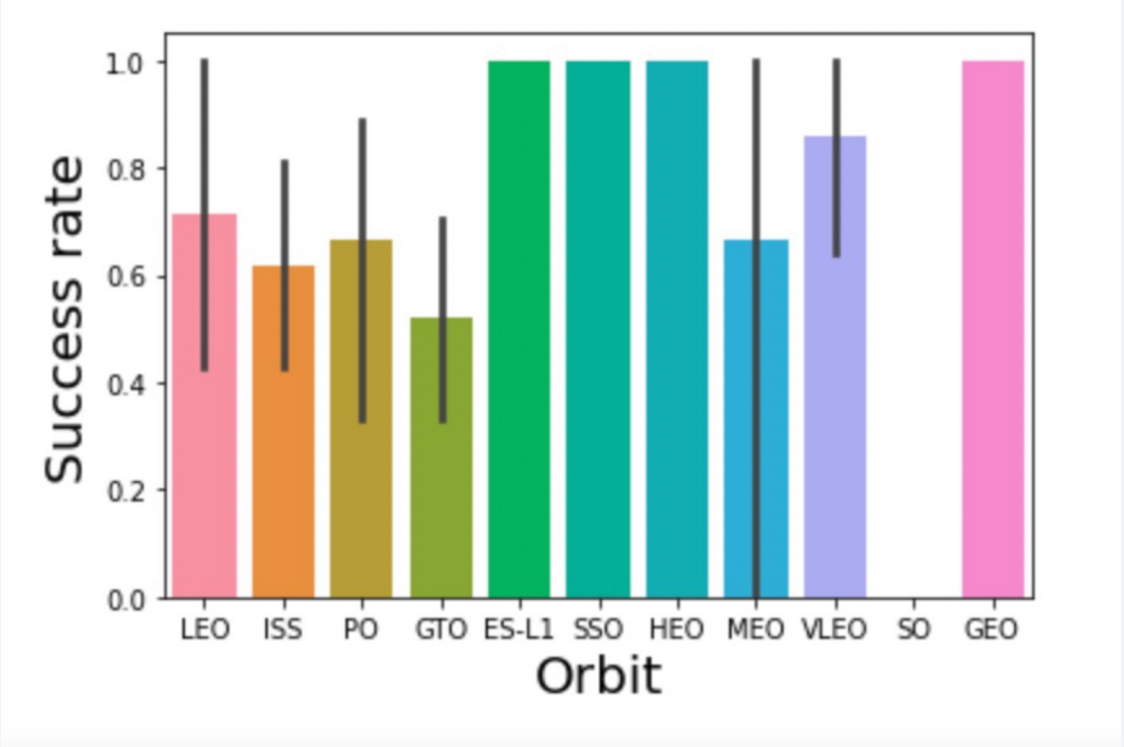
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site



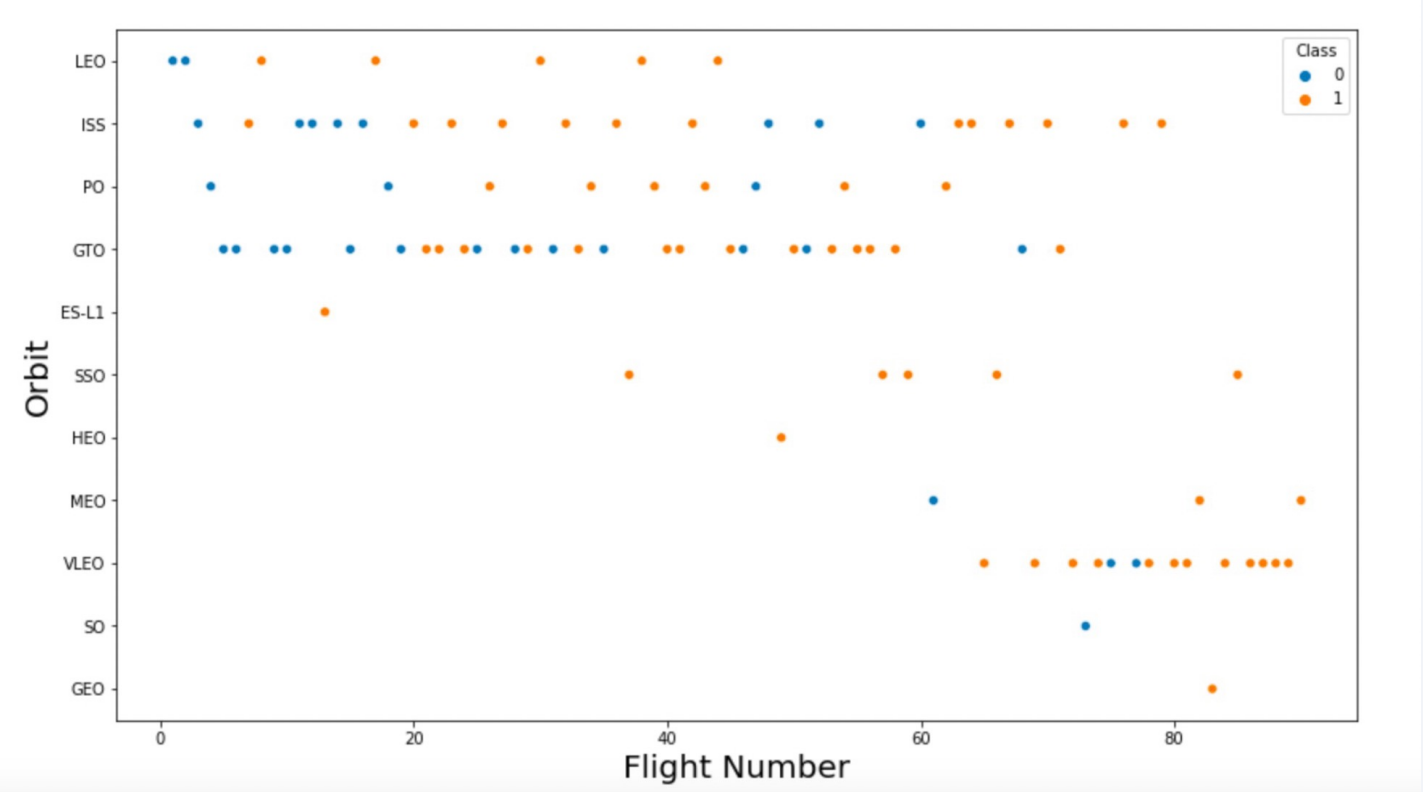
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type



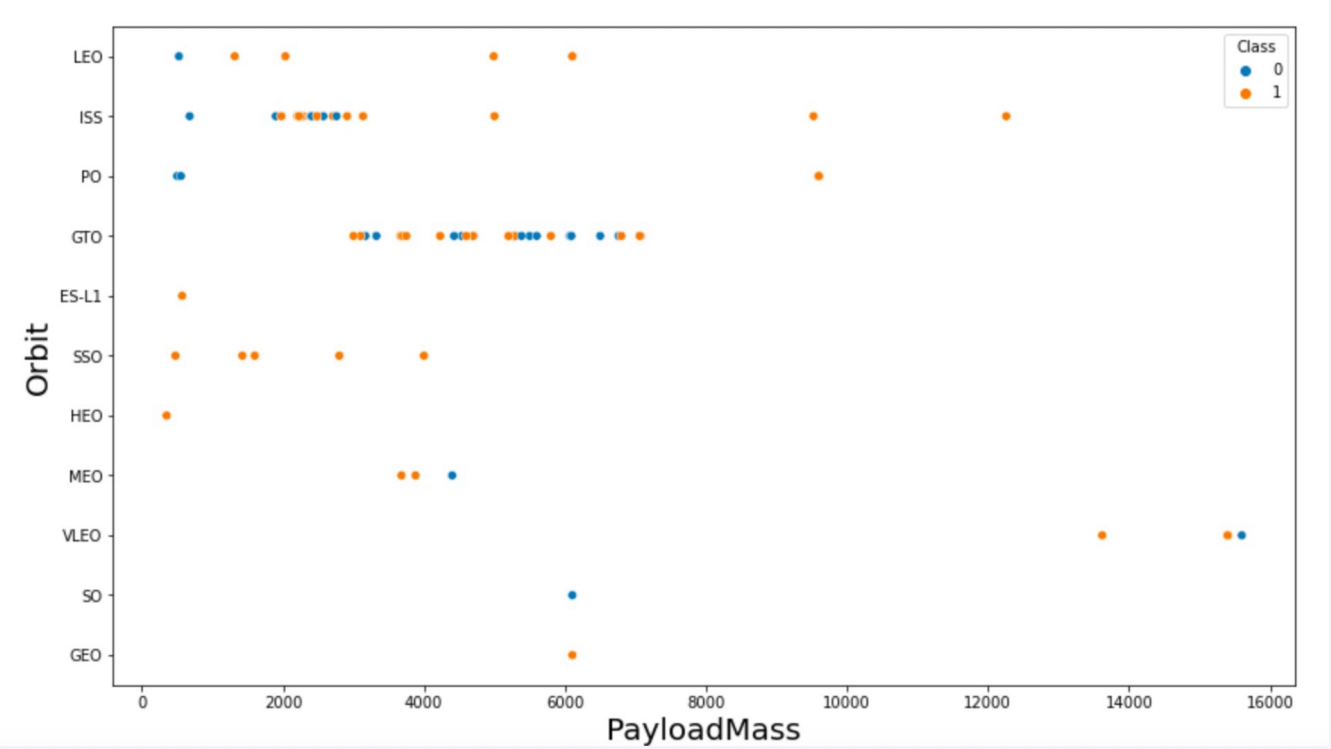
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type



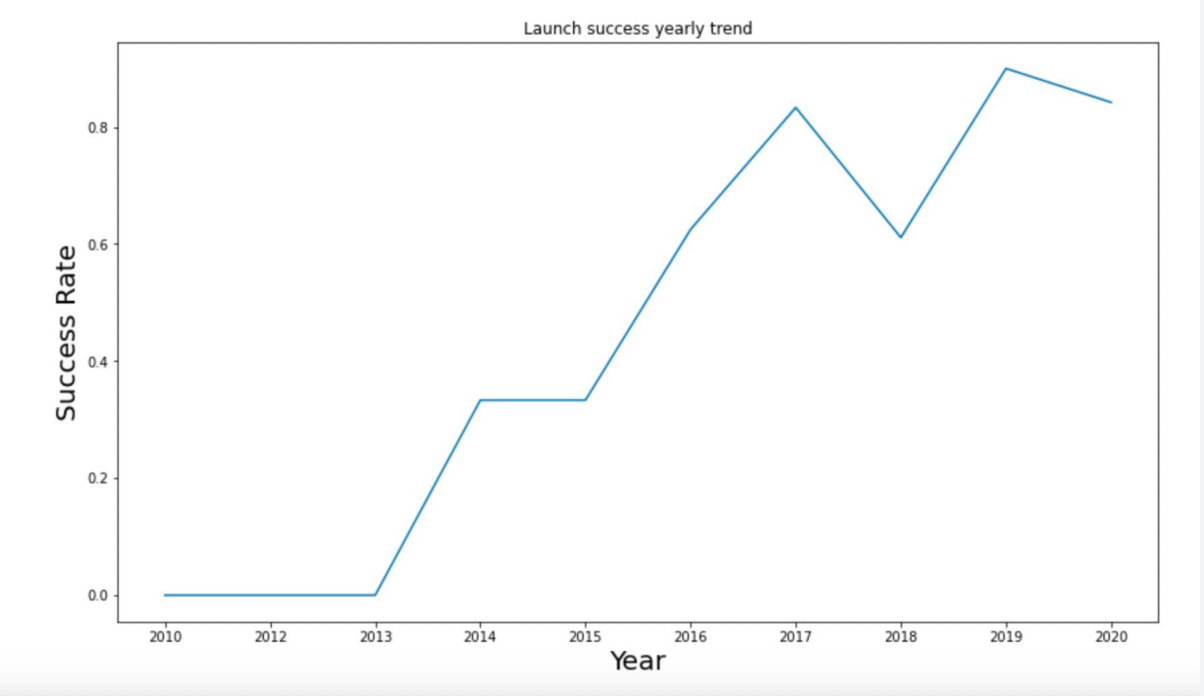
Payload vs. Orbit Type

- Scatter point of payload vs. orbit type



Launch Success Yearly Trend

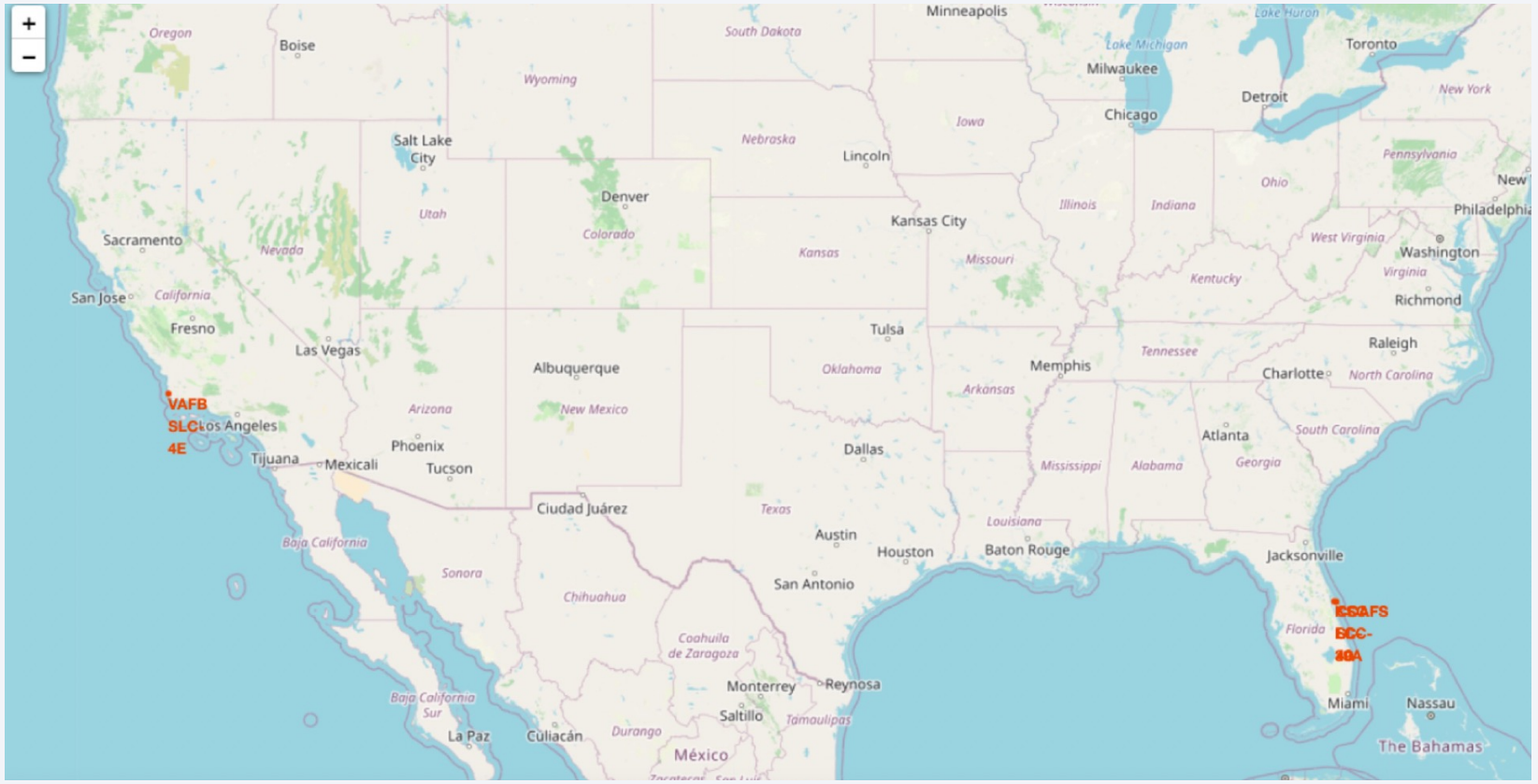
- Line chart of yearly average success rate




Section 3
Launch Sites
Proximities Analysis

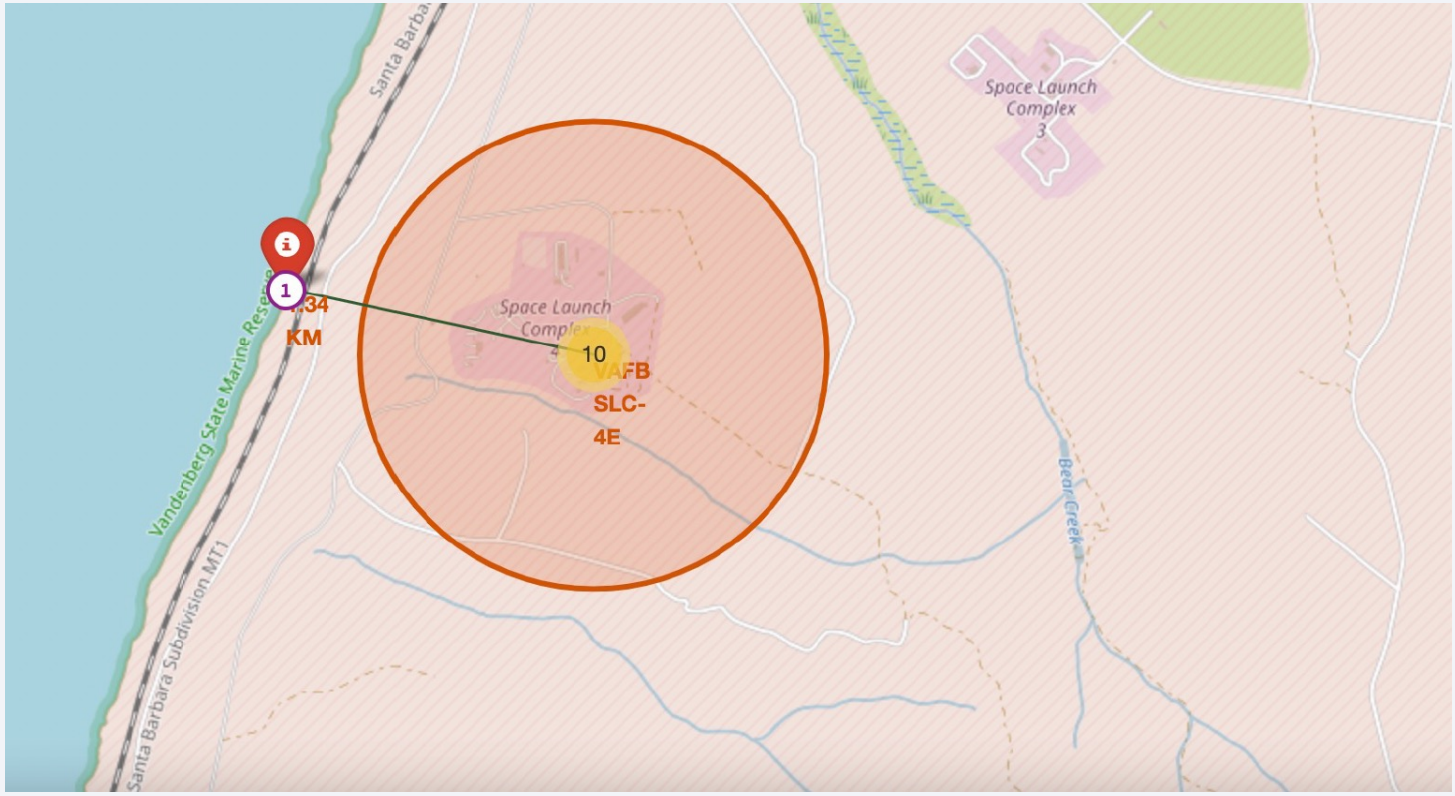
Folium

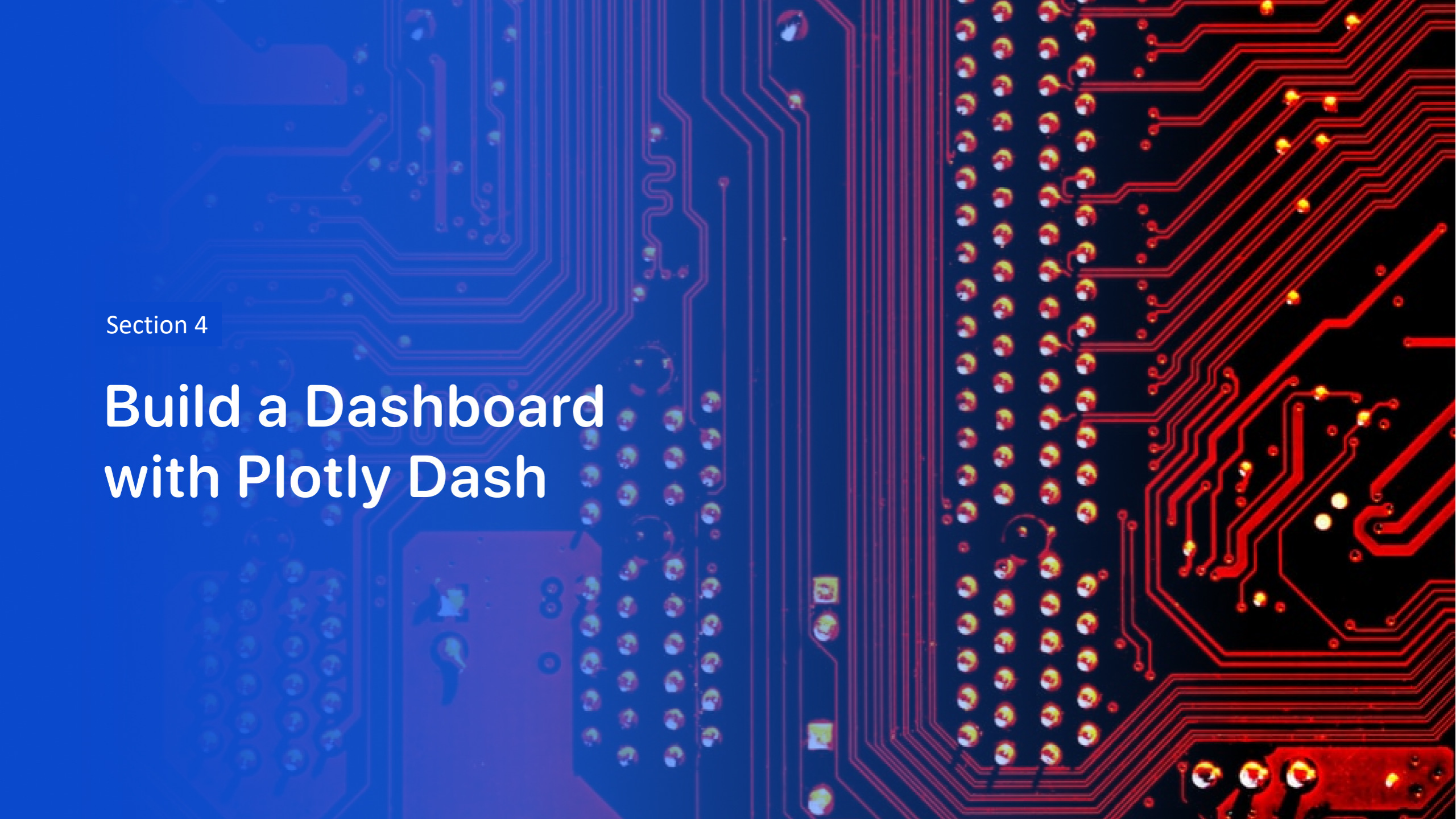
- Launch sites displayed on map



Folium

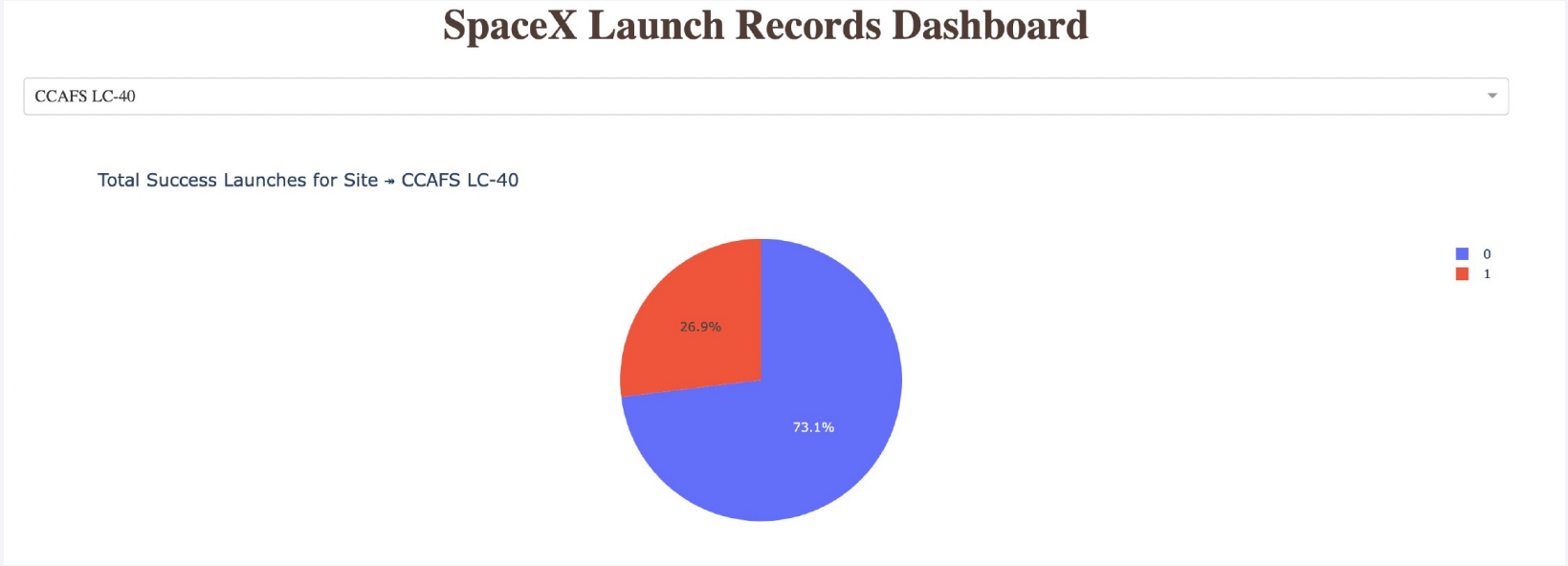
- Distance between launch sites to proximities such as nearest railway, city, or highway




Section 4
Build a Dashboard with Plotly Dash

Dashboard Screenshot 1

- Pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. 73.1% of launches done at CCAFS LC-40 are failed launches.



Dashboard Screenshot 2

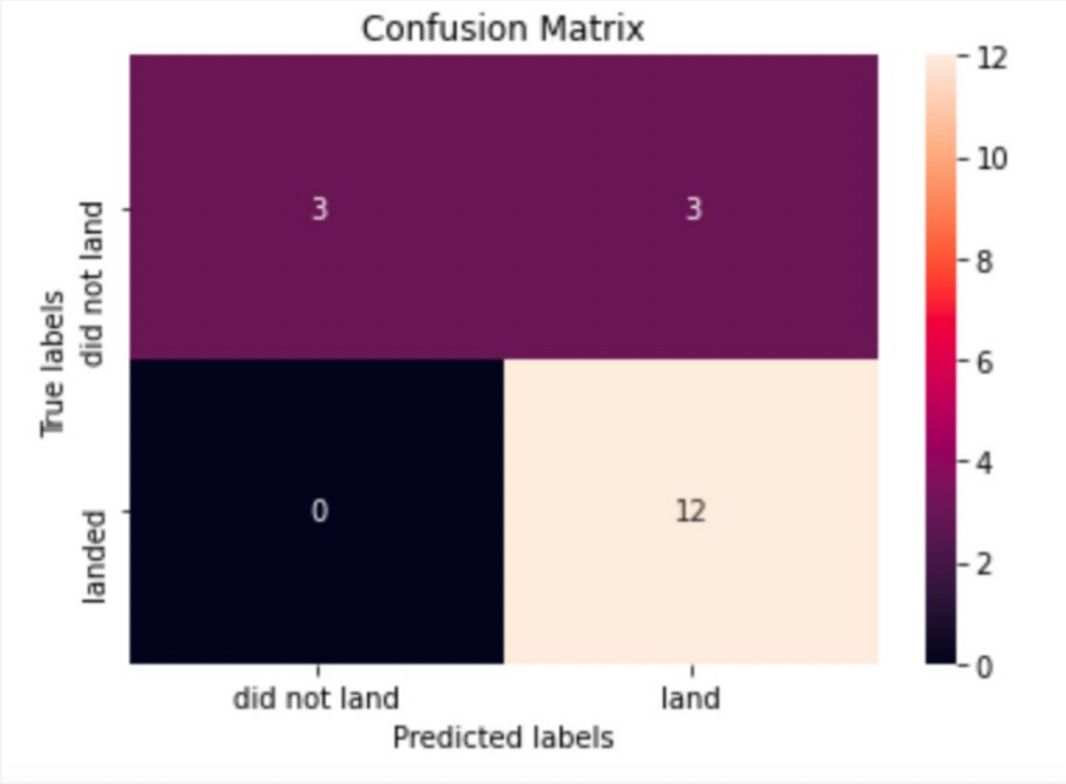
- Scatterplot shown where payload mass set to from 2000 to 8000 kg

Section 5

Predictive Analysis (Classification)

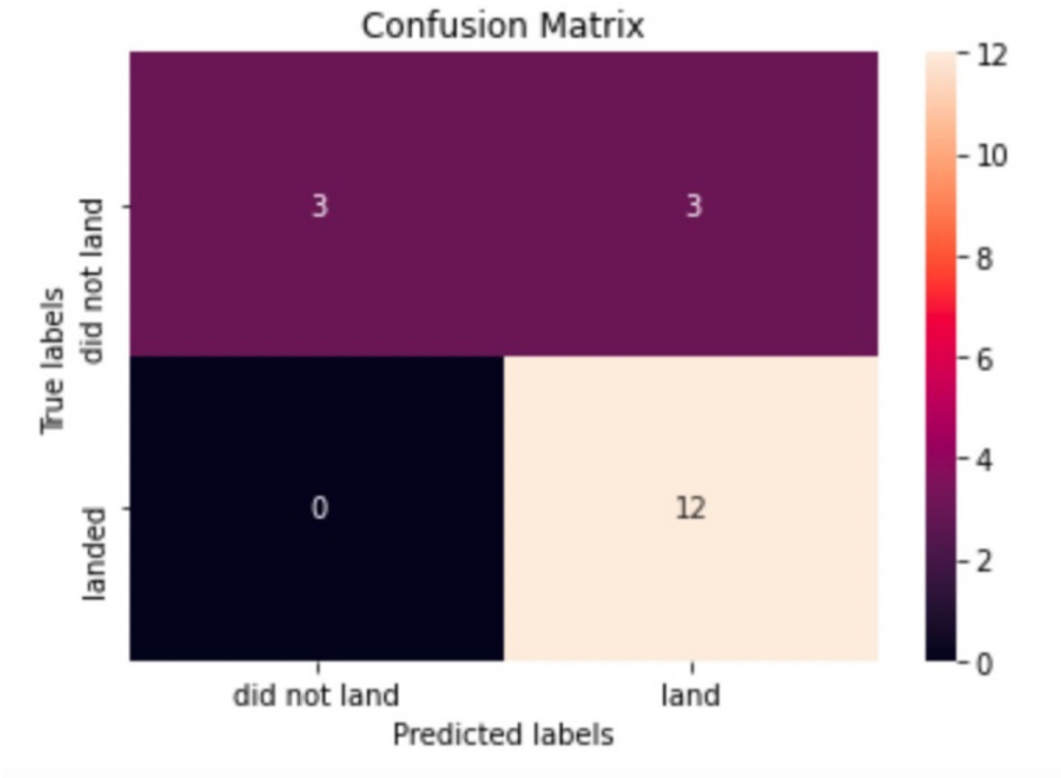
Classification Accuracy

- Logistic regression
  - GridSearchCV score: 0.846
  - Accuracy score on test set: 0.833
  - Confusion matrix:



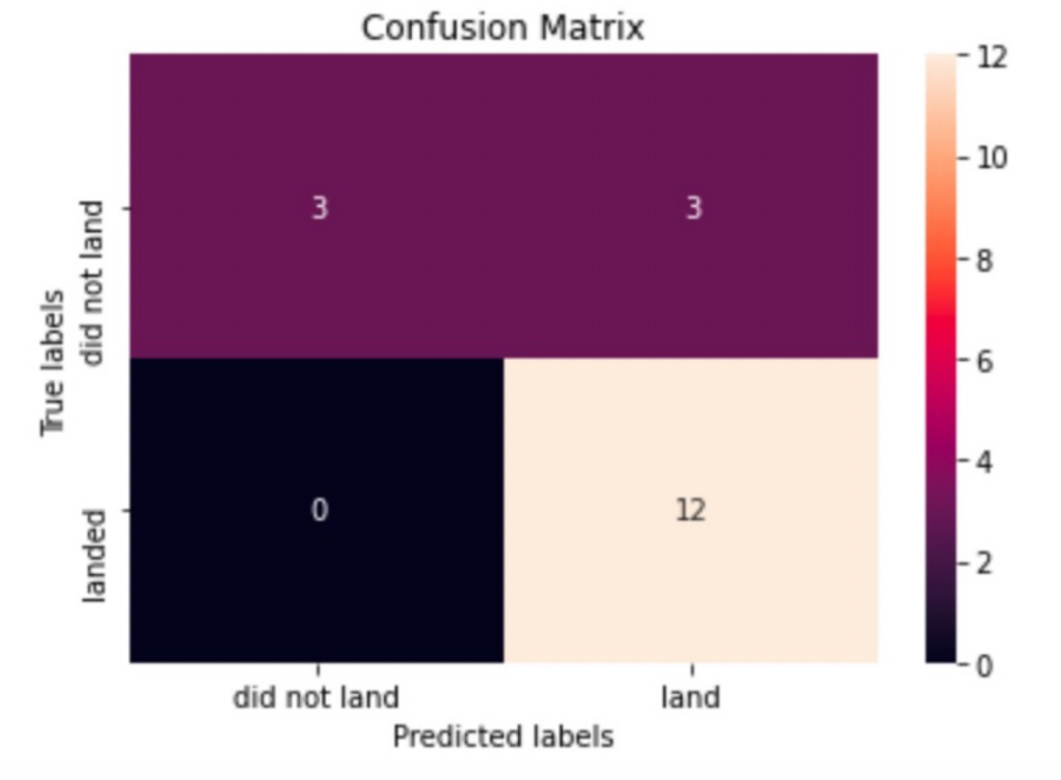
Classification Accuracy

- Support Vector Machine (SVM)
  - GridSearchCV score: 0.848
  - Accuracy score on test set: 0.833
  - Confusion matrix:



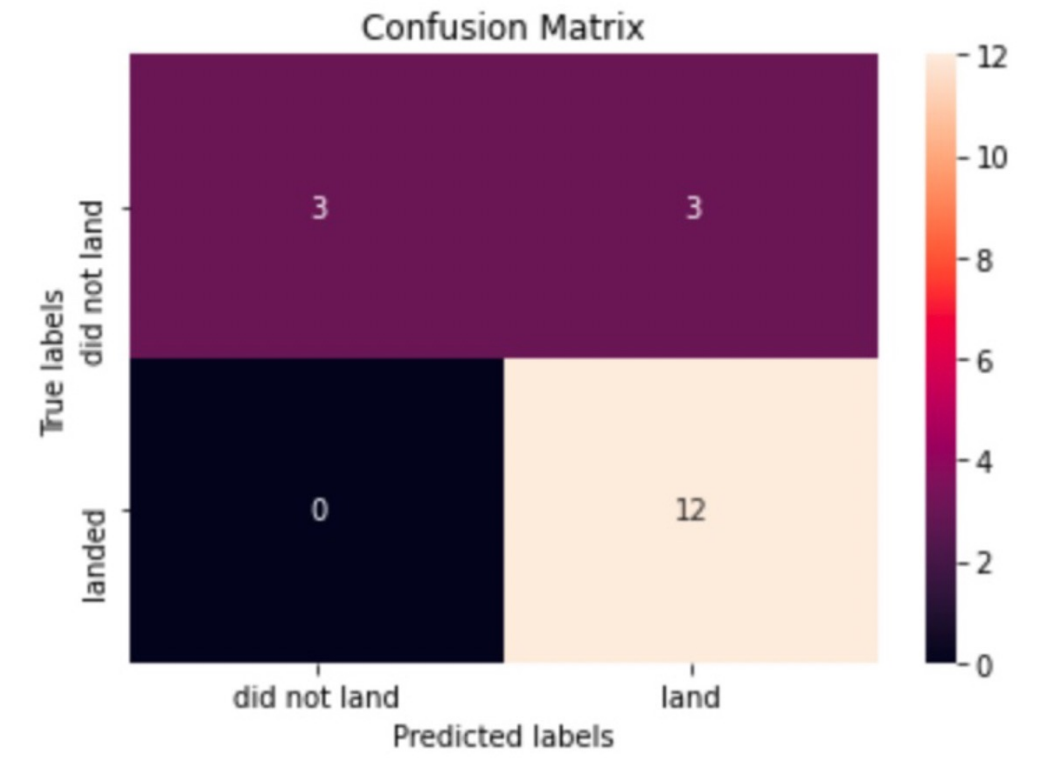
Classification Accuracy

- Decision Tree
  - GridSearchCV score: 0.889
  - Accuracy score on test set: 0.833
  - Confusion matrix:



Classification Accuracy

- KNN
  - GridSearchCV score: 0.847
  - Accuracy score on test set: 0.833
  - Confusion matrix:



Conclusions

- 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 decision tree algorithm performed the best among the 4 machine learning algorithms employed.

Thank you!