IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Anastasia Gracheva>
<10.04.2025>

Outline

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

Executive Summary

- Summary of methodologies

1. Data collection
2. Data wrangling
3. Data Analysis
4. Data Visualization
5. An interactive map with Folium
6. Predictive analysis

- Summary of all results

Introduction

- Project background and context

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.

- Problems to find answers to:

1. Determine the cost of a launch
2. Collect data on the Falcon 9 first-stage landings
3. Analyze the launch site proximity
4. Calculate distances on an interactive map
5. Determine if the first stage of Falcon 9 will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

1. API to extract information from a web service
2. Web scrapping

- Perform data wrangling

1. Convert landing outcomes into labels “1” (successfully landed) and “0” (unsuccessful landing)

- Perform exploratory data analysis (EDA) using visualization
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models

Data Collection

- API requests from SpaceX REST API such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Web Scraping data are Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

Request data from SpaceX API → Clean the requested data → Request and parse the SpaceX launch data using the GET request → Filter the dataframe to only include Falcon 9 launches → Calculate the mean for the PayloadMass → Replace missing values in the data with the mean

GitHub URL

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from its URL using an HTTP GET method → Extract all column/variable names from the HTML table header → Apply the "extract_column_from_header" to extract column name one by one → Create a data frame by parsing the launch HTML tables

GitHub URL

Data Wrangling

Load Space X dataset → Identify and calculate the percentage of the missing values in each attribute → Calculate the number of launches on each site using the method value_counts() → Calculate the number and occurrence of each orbit using the method .value_counts() → Calculate the number and occurence of mission outcome per orbit type using the method .value_counts() on the column Outcome → Create a landing outcome label from Outcome column

GitHub URL

EDA with Data Visualization

Read the SpaceX dataset into a Pandas dataframe and print its summary → Visualize the relationship between Flight Number and Launch Site using the function catplot → Visualize the relationship between Payload Mass and Launch Site → Visualize the relationship between success rate of each orbit type with a bar chart → Visualize the relationship between FlightNumber and Orbit type → Visualize the relationship between Payload Mass and Orbit type → Visualize the launch success yearly trend with a line chart with x axis to be Year and y axis to be average success rate → Create dummy variables to categorical columns using the function get_dummies and features dataframe → Cast all numeric columns to float64

GitHub URL

Build an Interactive Map with Folium

- Marked all launch sites on a map using their latitude and longitude coordinates
- Added markers of success (Green) and failed (Red) launches using Marker Cluster
- Added coloured Lines to show distances between Launch Sites and their proximities to Railway, Highway, Coastline and Cities near them

GitHub URL

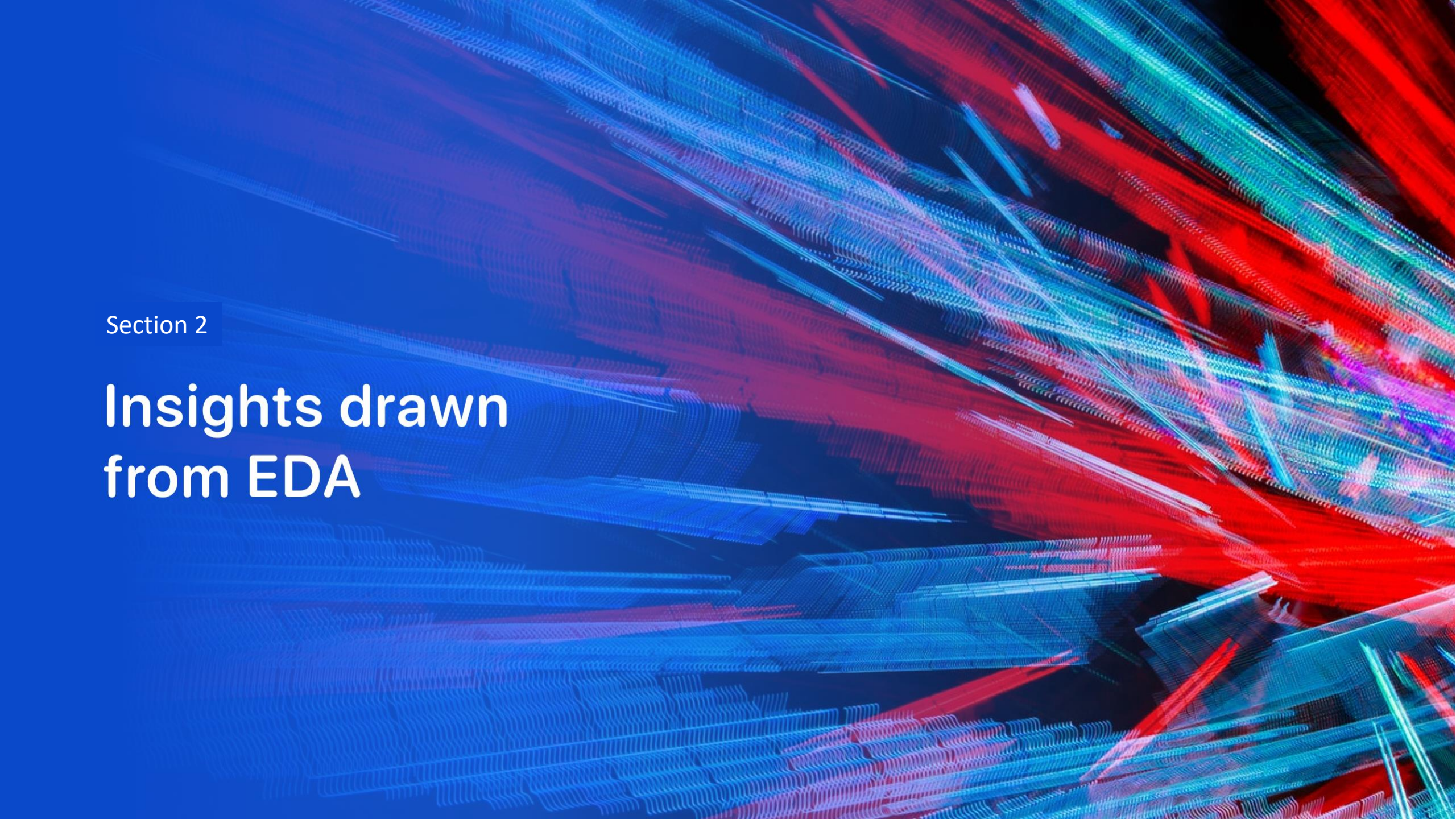
Predictive Analysis (Classification)

- Create a NumPy array from the column Class in data, by applying the method to_numpy() then assign it to the variable Y → Standardize the data in X then reassign it to the variable X → Use the function train_test_split to split the data X and Y into training and test data → Create a logistic regression object then create a GridSearchCV object logreg_cv → Calculate the accuracy on the test data using the method score → Create a support vector machine object then create a GridSearchCV object svm_cv → Calculate the accuracy on the test data using the method score → Create a decision tree classifier object then create a GridSearchCV object tree_cv → Calculate the accuracy of tree_cv on the test data using the method score → Create a k nearest neighbors object then create a GridSearchCV object knn_cv → Calculate the accuracy of knn_cv on the test data using the method score → Find the method performs best

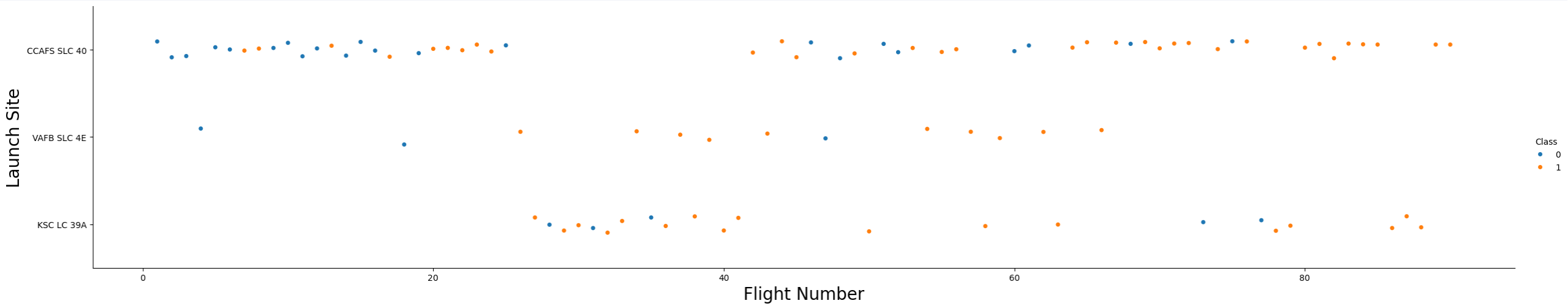
GitHub URL

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

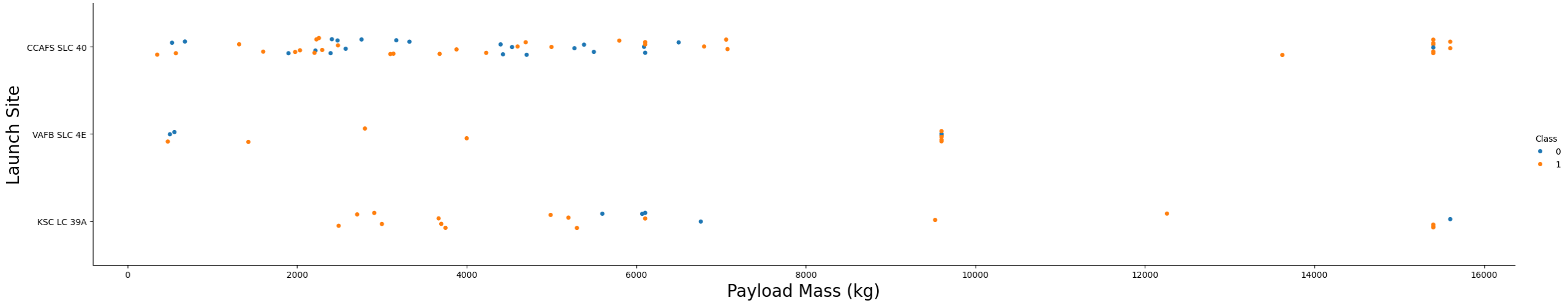

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



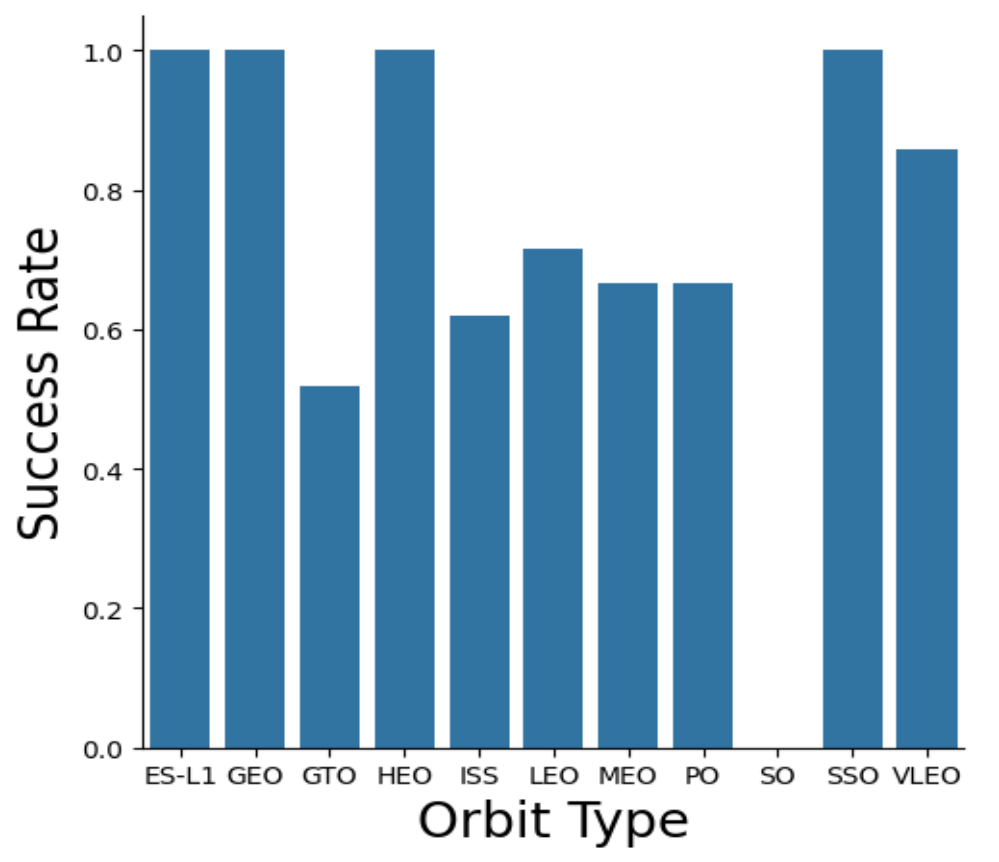
- The CCAFS SLC 40 launch site has almost a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.

Payload vs. Launch Site



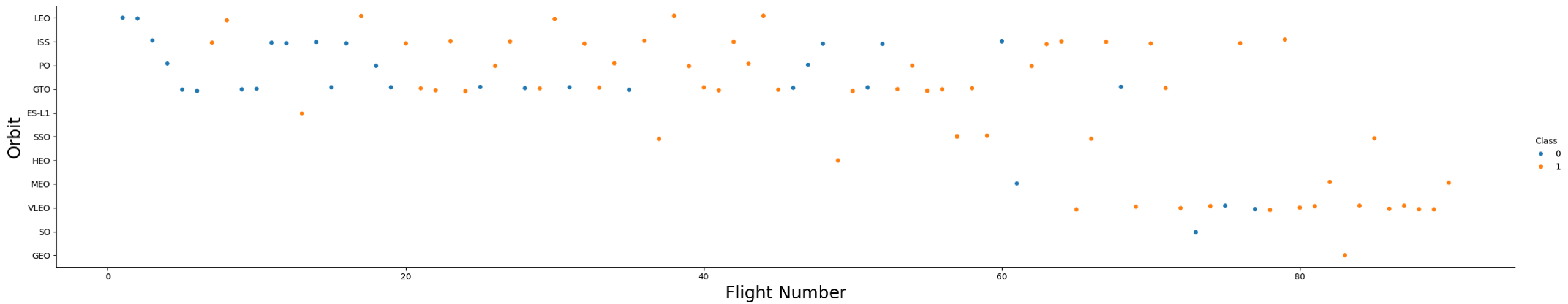
- The higher the payload mass, the higher the success rate.
- However, KSC LC 39A has a 100% success for smaller payload mass.

Success Rate vs. Orbit Type



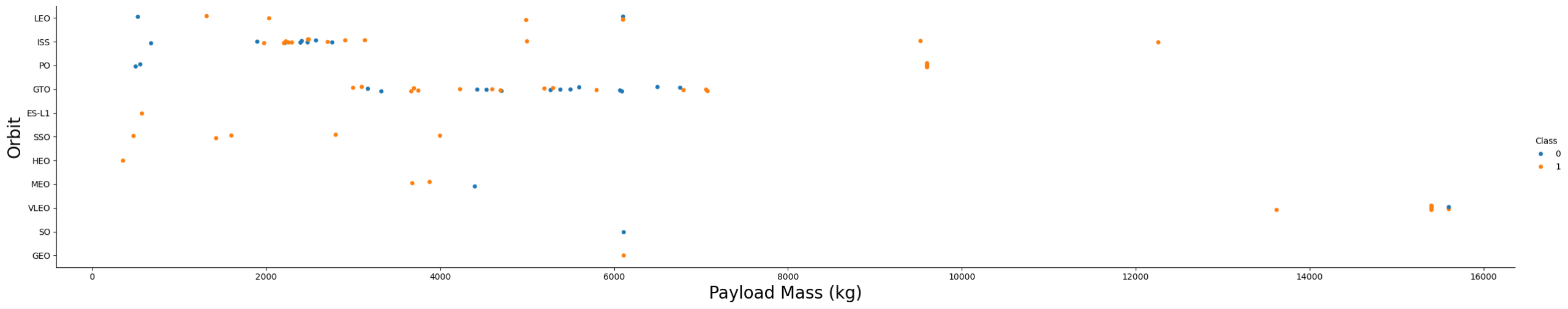
- Orbits with 100% success: ES-L1, GEO, HEO, SSO.
- Orbits with 0% success: SO.
- Orbits with success between 50% and 85%: GTO, ISS, LEO, MEO, PO, VLEO.

Flight Number vs. Orbit Type



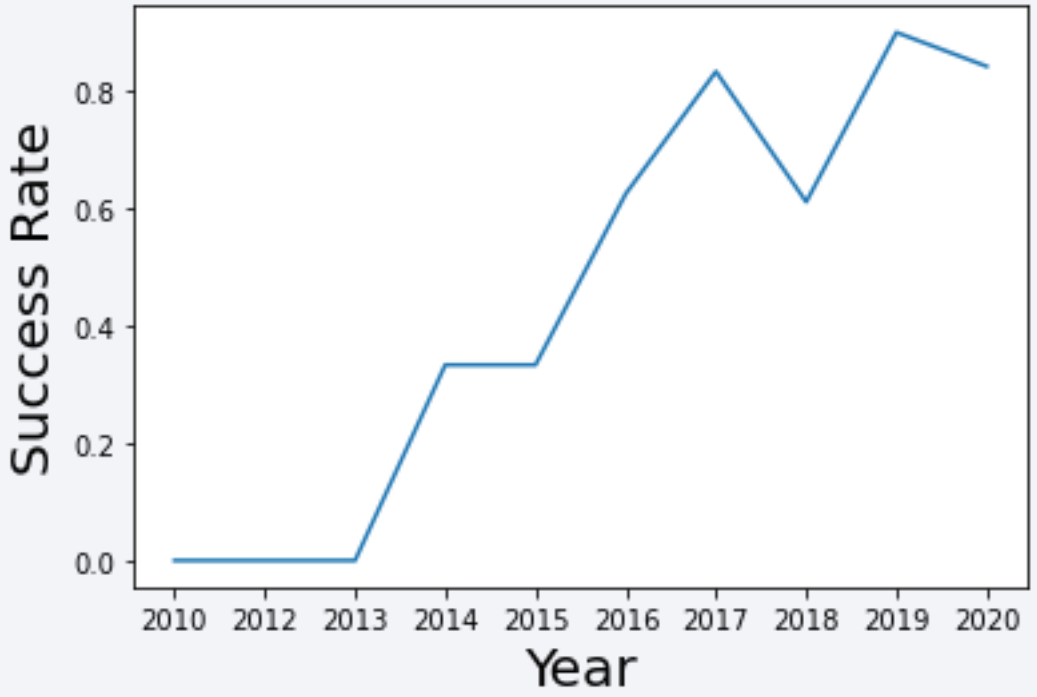
- Success is related to the number of flights in the LEO orbit.
- Success is not related to the number of flights in GTO orbit.

Payload vs. Orbit Type



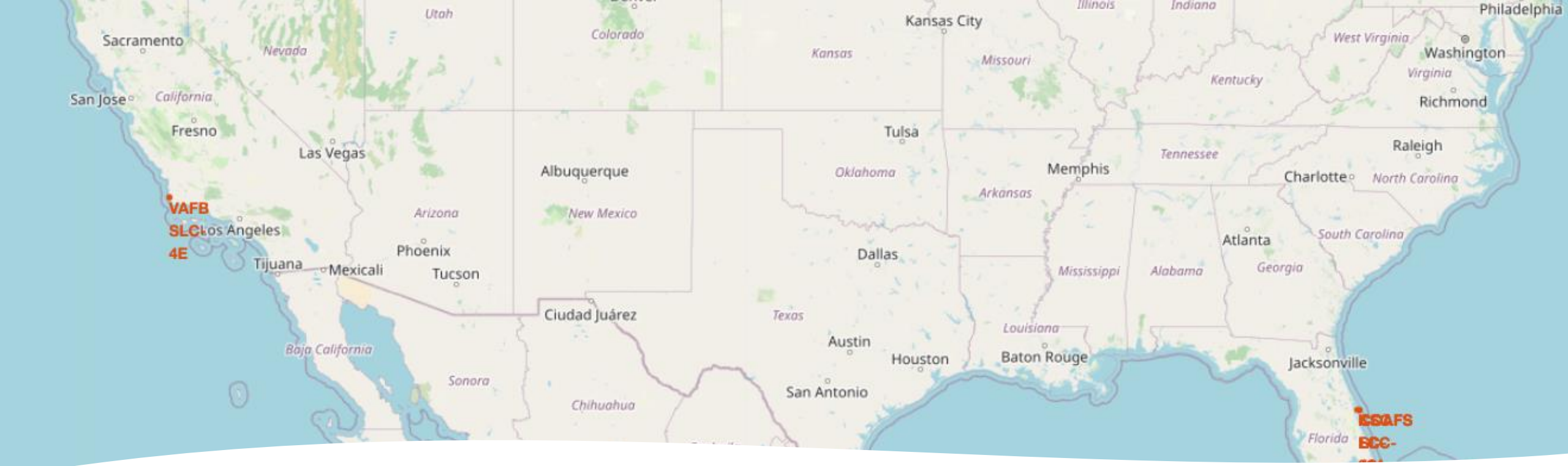
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



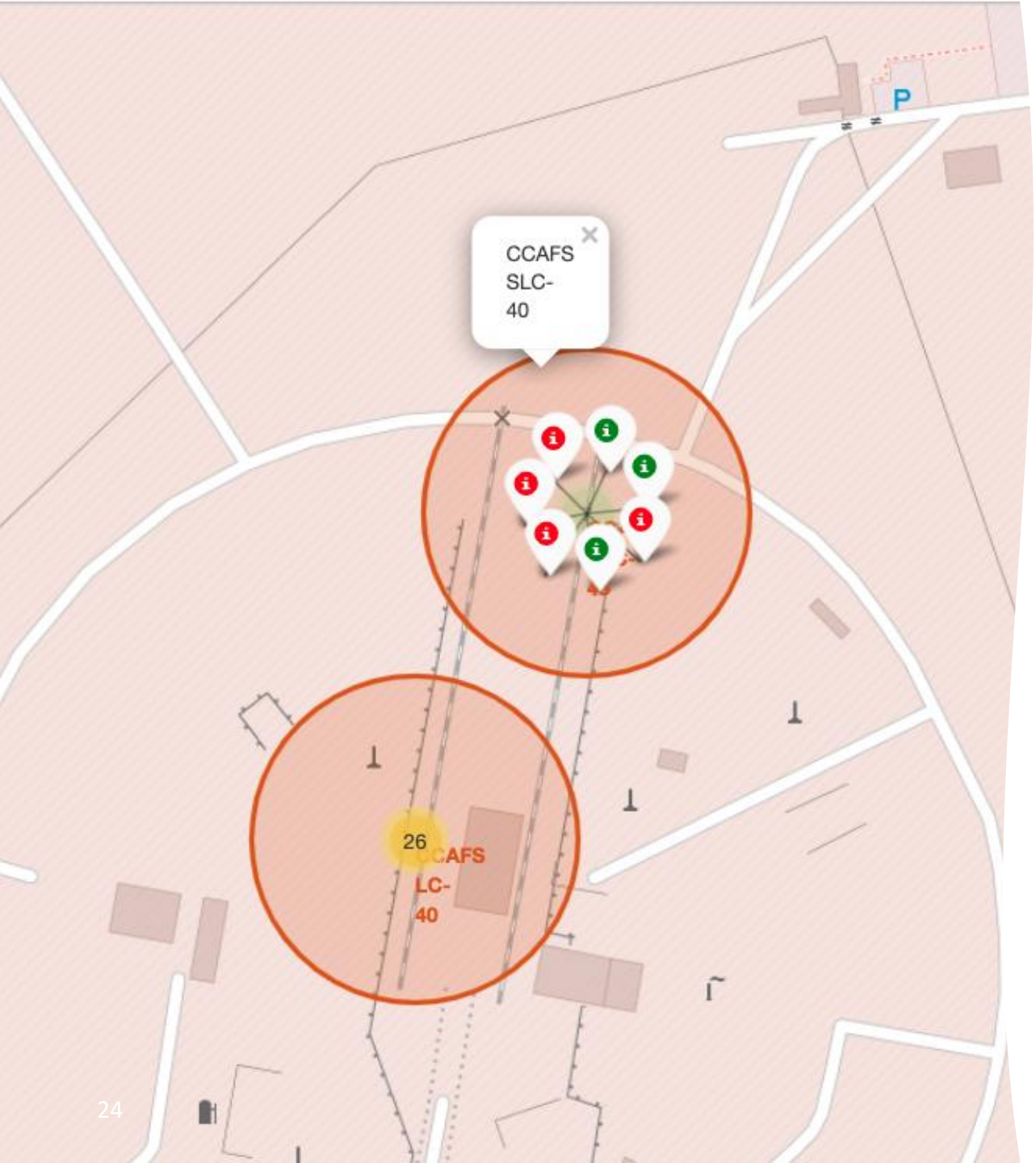
- The sucess rate since 2013 kept increasing till 2020.


Section 3
Launch Sites
Proximities Analysis



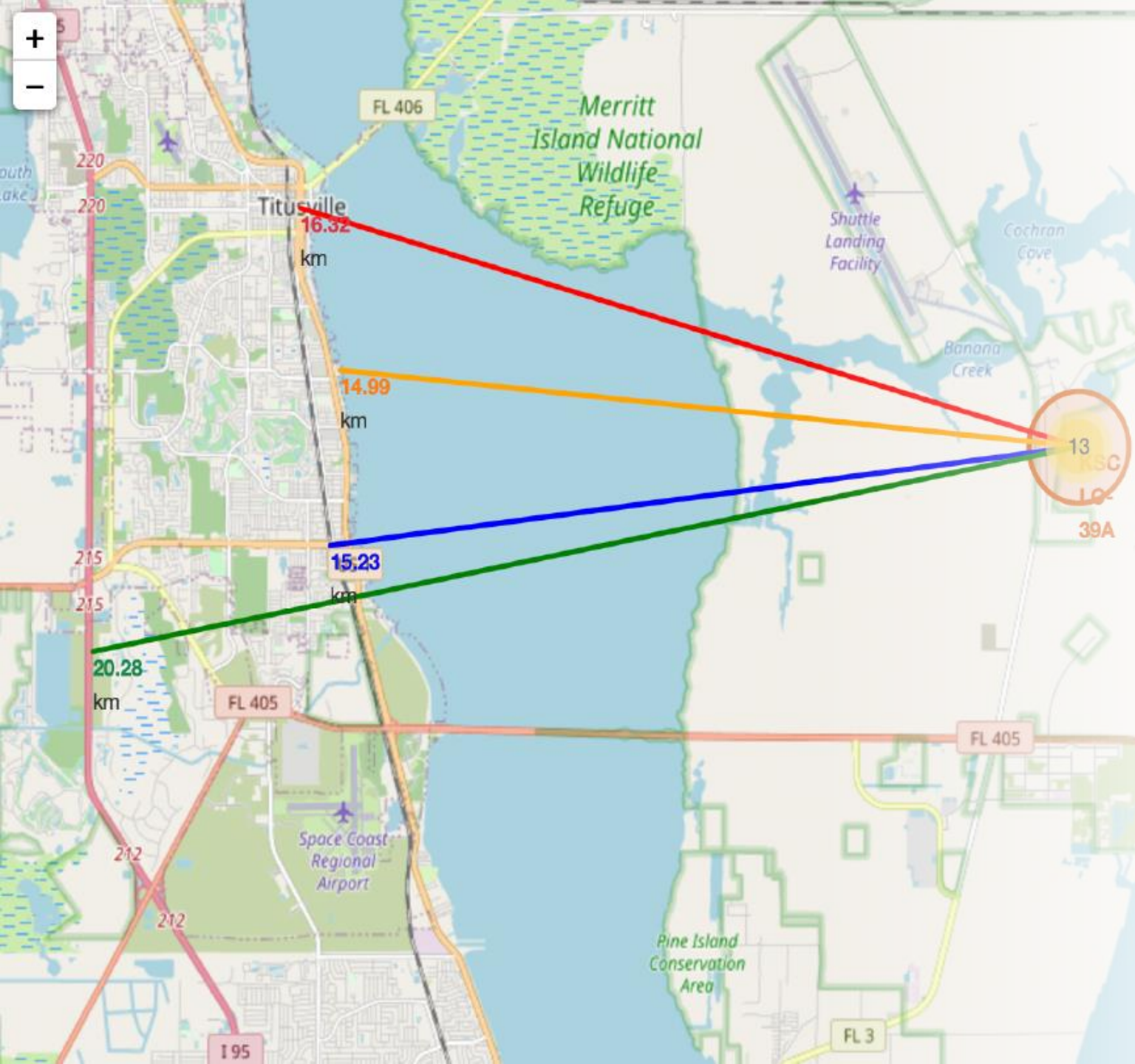
All marked launch sites

- Most of launch sites are near the Equator to make rockets move faster.
- All launch sites are near the coast to minimise the danger to people.



Color-labeled markers

- Green Marker means successful launch
- Red Marker means failed launch



Launch site KSC LC-39A and its proximities

- KSC LC-39A is close to railway (15.23 km), close to highway (20.28 km), close to the coastline (14.99 km), close to the nearest city (16.32 km)
- The site is potentially dangerous to nearby population

Section 5

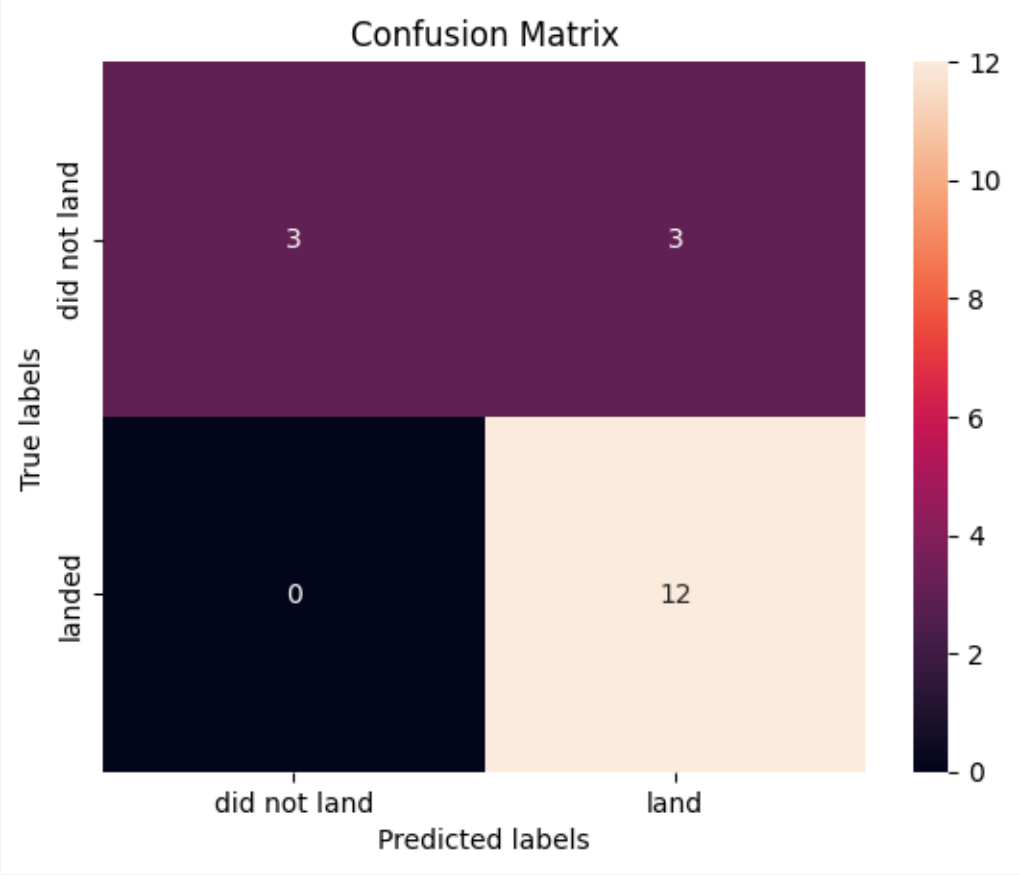
Predictive Analysis (Classification)

Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.666667	0.819444
F1_Score	0.909091	0.916031	0.800000	0.900763
Accuracy	0.866667	0.877778	0.666667	0.855556

- The best model is SVM model.

Confusion Matrix



- SVM accuracy is 0.8333333333333334.

Conclusions

- SVM model performs best
- Launch sites are near the highways to transport personal and light cargo.
- Launch sites are near the railways to transport heavy cargo.
- Most of the launch sites are not near cities to minimizes safety risks for people.
- The success rate of launches increases over the years.
- KSC LC 39A has the highest success rate.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Thank you!