IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Walmond3
11 Oct 2024

Outline

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

Executive Summary

- This project uses the SpaceX API and web scrapping from Wikipedia to get the data through data wrangling and formatting to clean the data into a format that is more suitable for modelling.
- Various EDA techniques are used to visualized the data, like scatter plot, bar chart, line chart, building map and dashboard.
- 4 machine learning models are used to predict the landing outcome of the launches, and further tuned using GridSearchCV
- There are 4 launch sites in total, and the success rate of launch increases throughout the year except 2018
- Orbit types that have highest success rate are ES-L1, GEO, HEO and SSO
- The best model is Decision Tree after tuning, contributing the highest accuracy.

Introduction

- This project is to predict if the Falcon 9 first stage will land successfully. If the first stage lands successfully, we can reuse it and determine the cost of launching a rocket. Thus, we can minimize the rocket launch cost.
- Problems:

1. What are the relationship between landing outcomes with other features?
2. Which orbit has the highest success rate?
3. Which launch site has the highest success rate?
4. Which model has the highest accuracy?

Section 1

Methodology

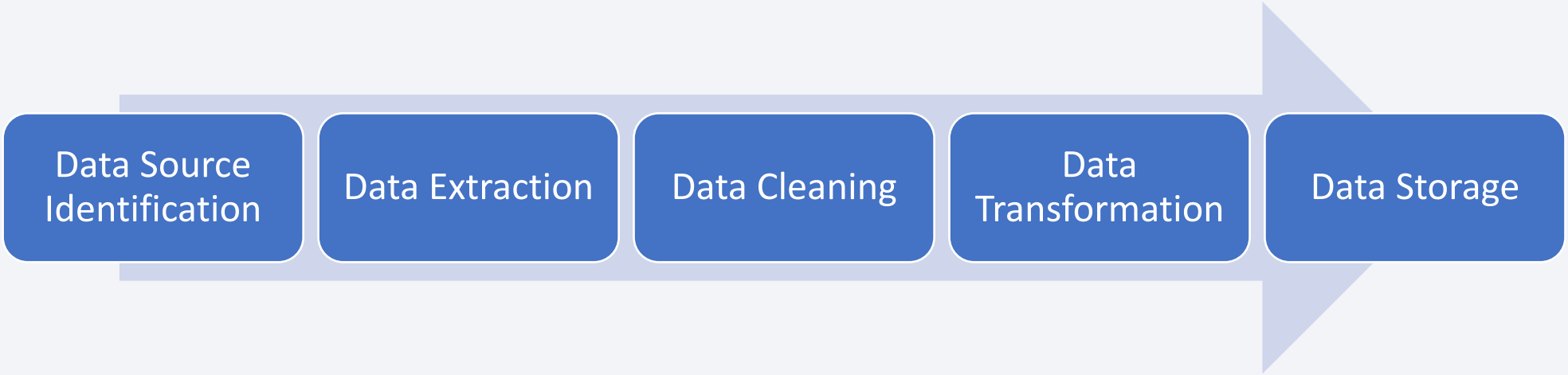
Methodology

Executive Summary

- Data collection methodology:
  - Through API and web scraping from Wikipedia
- Perform data wrangling
  - Perform EDA on the data and determine the training labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Model: Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbour
  - Hyperparameter tuning: GridSearchCV
  - Evaluation: accuracy, classification matrix

Data Collection

- The datasets were collected through calling to SpaceX API and also web scrapping on Wikipedia
- The datasets were cleaned before converting into csv format

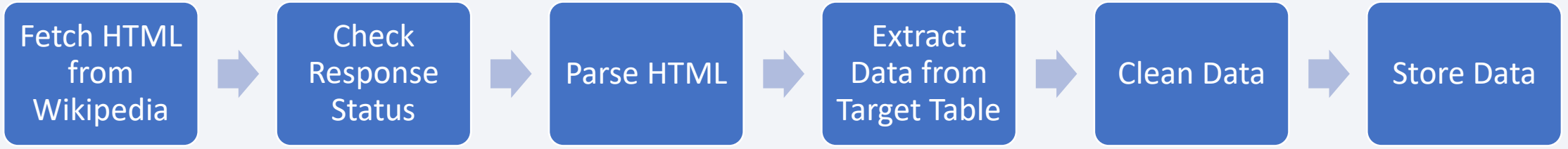


Data Collection – SpaceX API

- GitHub URL

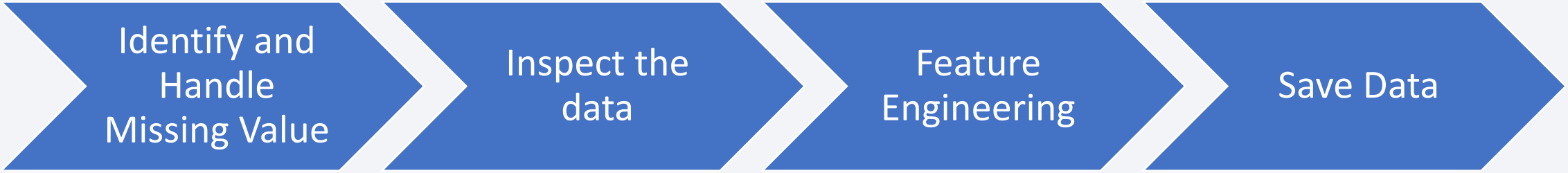
Data Collection - Scraping

- GitHub URL



Data Wrangling

- Identify and handle missing values
- Calculate number of launches on each site
- Calculate number and occurrence of each orbit, also with the outcomes
- Feature engineering: create new column, landing outcome label
- GitHub URL



EDA with Data Visualization

- Charts plotted: scatter point chart, bar chart, line chart
- Scatter point chart: To visualize the relationship between

1. Flight Number and Launch Site
2. Payload Mass and Launch Site
3. Flight Number and Orbit Type
4. Payload Mass and Orbit Type

- Bar chart: To visualize relationship between orbit type and success rate
- Line chart: To visualize the yearly trend of launch success
- GitHub URL

EDA with SQL

- Create SpaceX table using CREATE TABLE
- Find the unique launch sites using DISTINCT
- Calculate the total payload mass using SUM, and its average using AVG
- Find the first successful landing date using MIN
- Find the booster version which have success in drone ship and have payload mass greater than 4000 but less than 6000 using WHERE and BETWEEN
- Find the total successful and failure mission using CASE and GROUPBY
- Use ORDER to rank the result and LIMIT to restrict the number of result showed
- GitHub URL

Build an Interactive Map with Folium

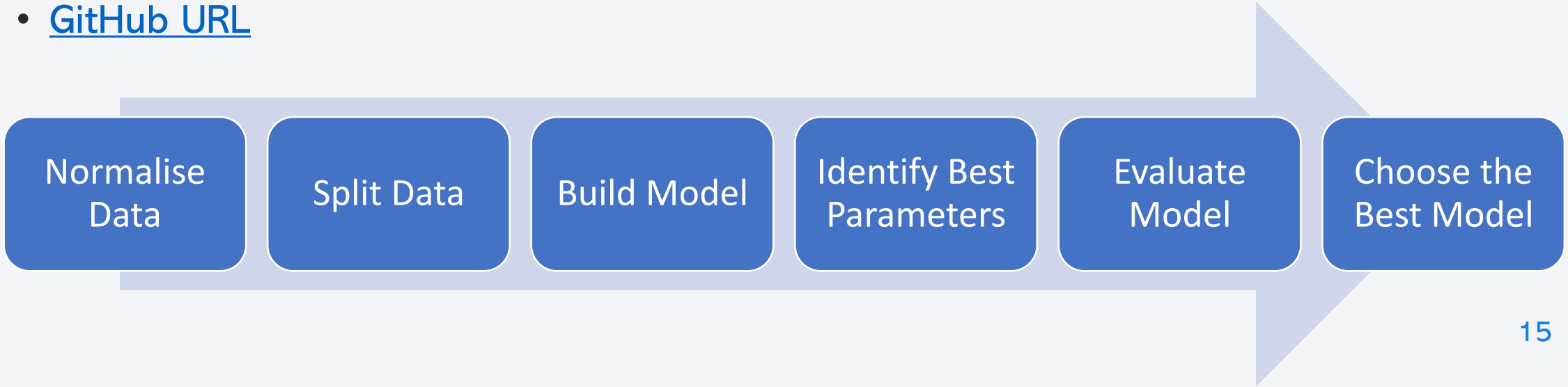
- Circle and Marker: to mark all launch sites on the map
- MarkerCluster: to cluster the success and failed launches of each launch site
- Marker and Polyline: to draw the line between launch site and coastline / railway / highway point, and denoted the distance between 2 points
- GitHub URL

Build a Dashboard with Plotly Dash

- Dropdown: To select the launch site
- Pie chart: To display the success rate of each launch site
- Slider: To select the payload range
- Scatter plot: To display the relationship between the selected payload range and the launch site
- GitHub URL

Predictive Analysis (Classification)

- Split the data into training and testing set
- Find the best parameters using GridSearchCV
- Evaluation through accuracy score and classification matrix
- Best model: highest accuracy
- GitHub URL

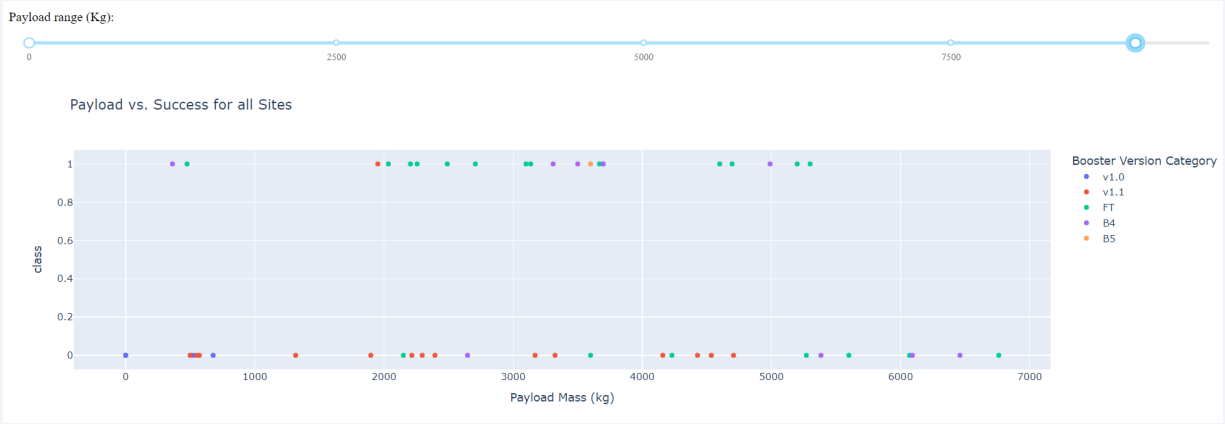


Results

- Exploratory data analysis results:

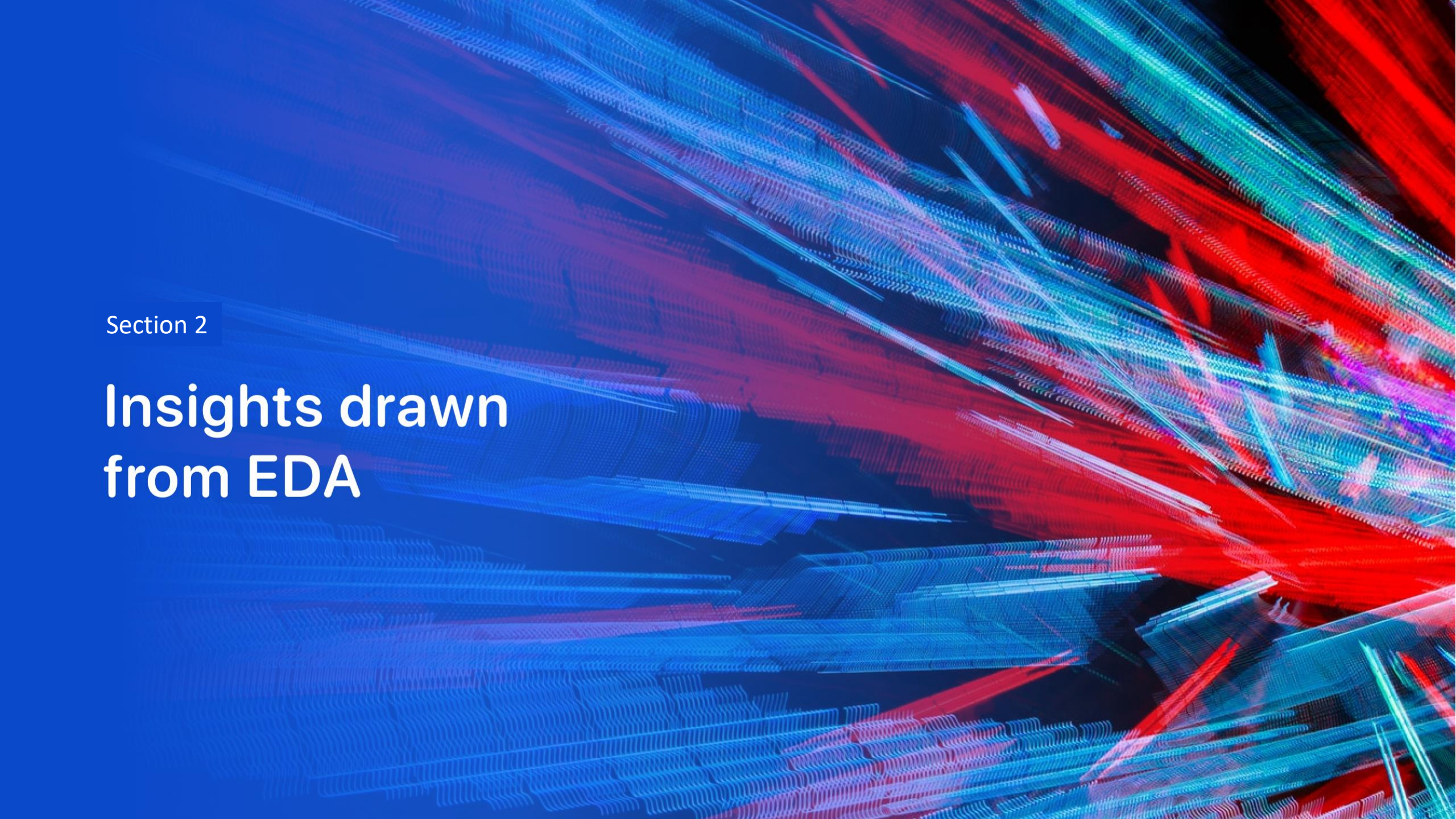
1. Yearly trend of success rate increases
2. ES-L1, GEO, HEO and SSO have the highest success rate
3. There are 4 different launch site
4. The most landing outcome is not attempt

- Interactive analytics demo:

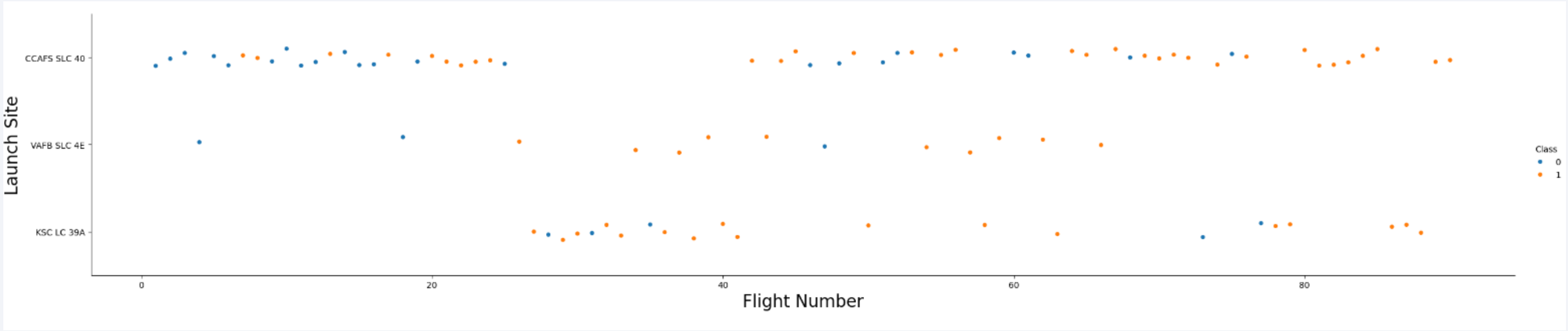


- Predictive analysis results:

1. Decision Tree is the best model (highest accuracy)

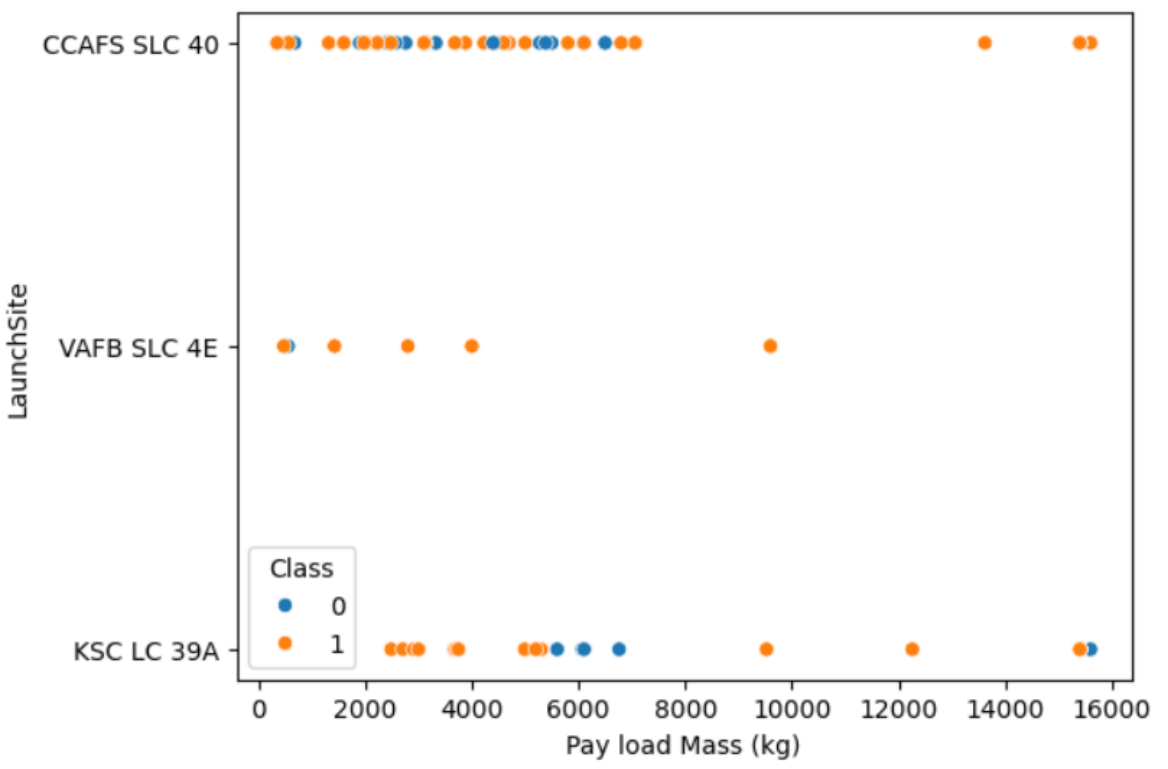

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



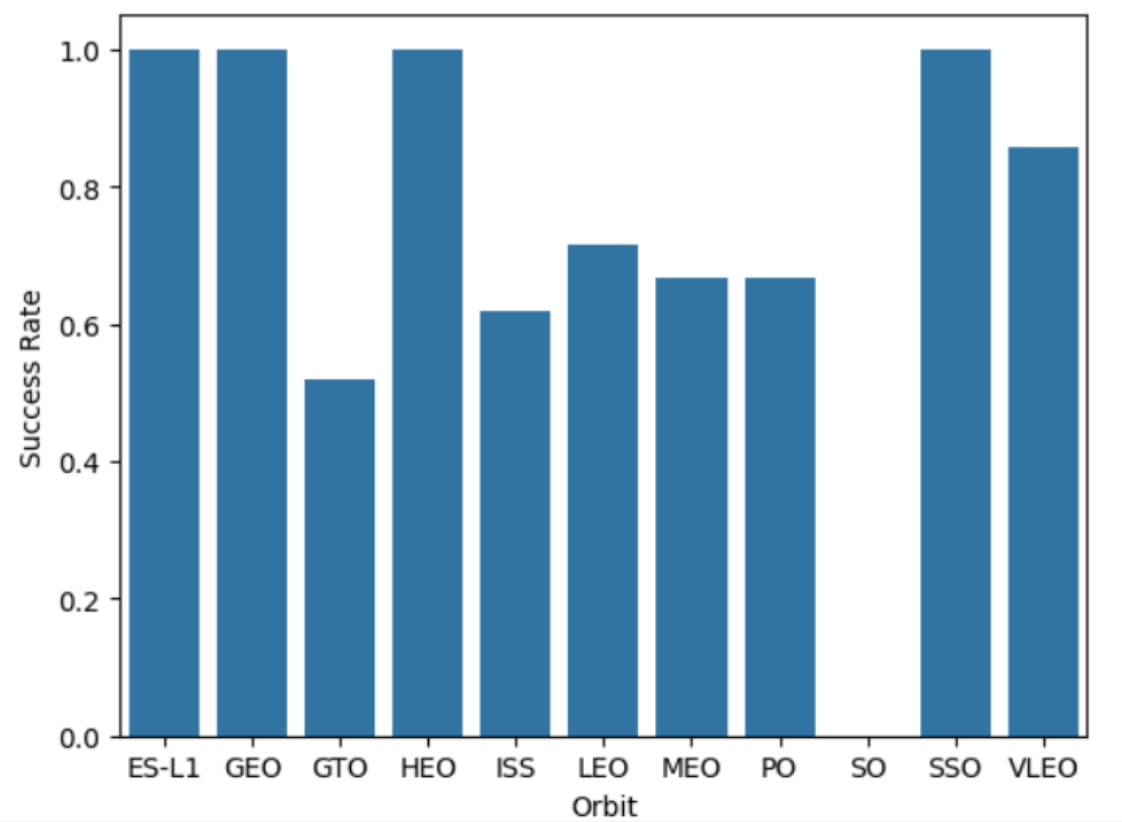
- The flight number increases, success outcomes for CCAFS SLC 40 increases
- More flight number, success outcome increases
- KSC LC 39A flight number focuses between 25 and 40
- VAFB SLC 4E flight number focuses between 30 and 70

Payload vs. Launch Site



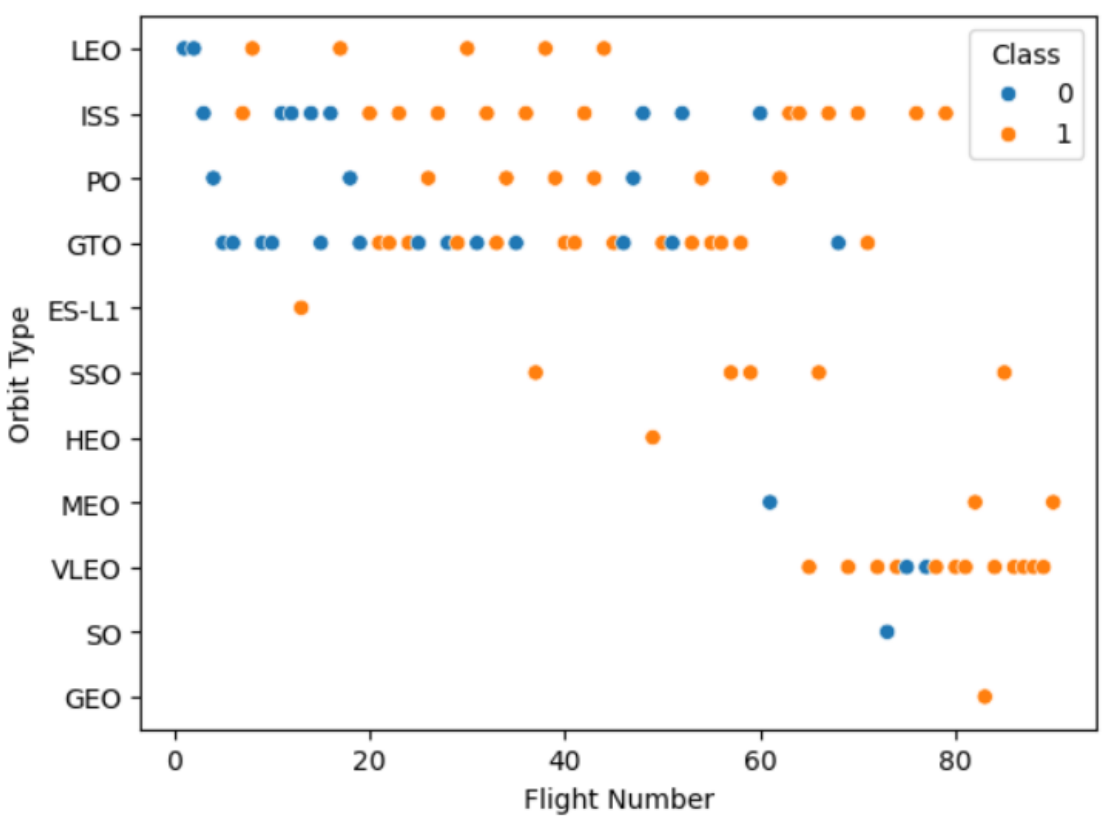
- VAFB-SLC no rockets launched for heavy payload mass (greater than 10000)
- More rockets (less than 8000 kg) are launched for CCAFS SLC 40
- KSC LC 39A has rockets launched for different payload mass

Success Rate vs. Orbit Type



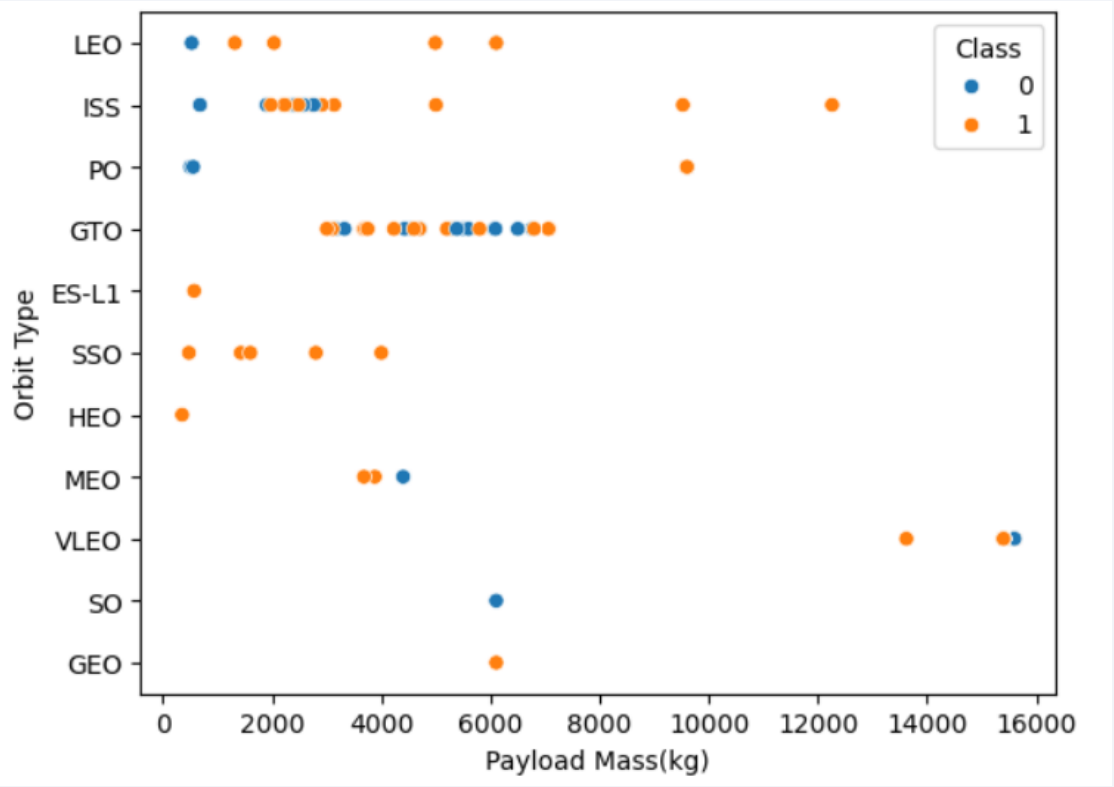
- ES-L1, GEO, HEO and SSO have the highest success rate
- SO has the lowest success (no success)
- VLEO has the second highest success rate (more than 0.8)

Flight Number vs. Orbit Type



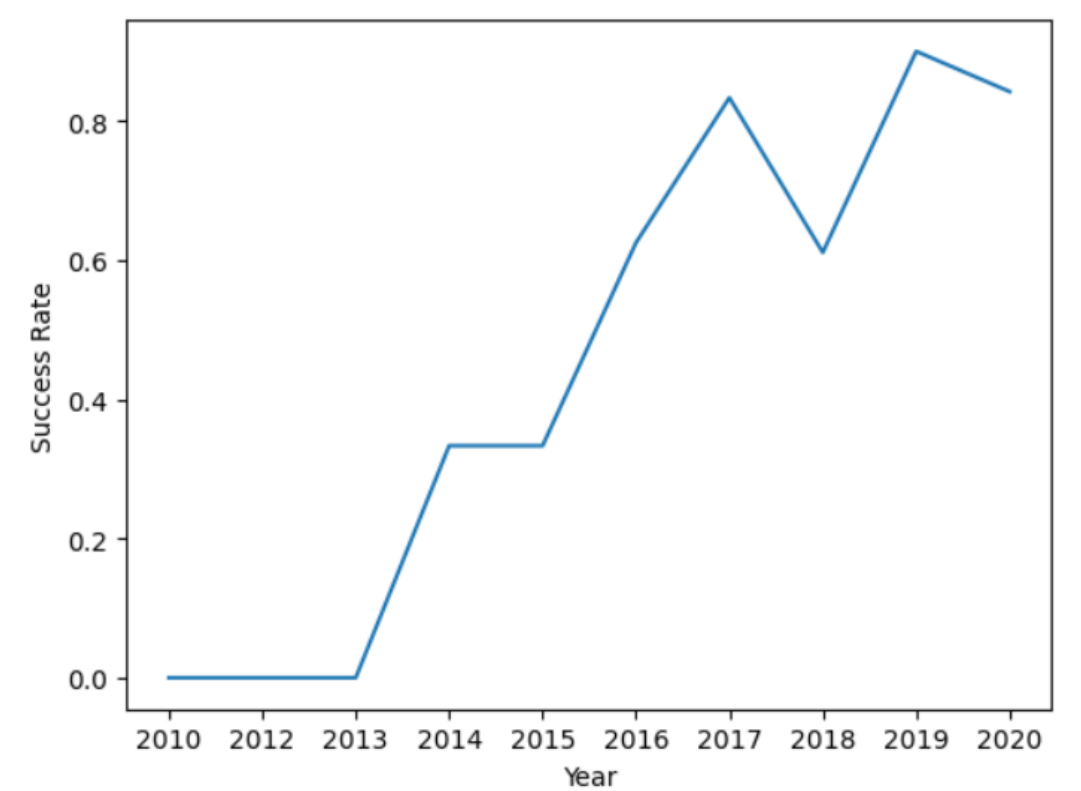
- ES-L1, SSO, HEO, GEO have all the launches succeeded
- ISS have the success launches between 20 and 40, and more than 60
- LEO, VLEO success rate increases as the flight number grows
- GTO does not show a specific relationship between flight number and orbit type

Payload vs. Orbit Type



- SSO has no payload mass greater than 5000kg
- VLEO has no payload mass less than 13000kg
- HEO payload mass are focusing between 3000kg and 5000kg
- GTO payload mass are focusing between 2000kg and 8000kg
- ISS success rate increase as payload mass increases

Launch Success Yearly Trend



- The success increases from 2013 till 2017
- There is a sudden drop in success rate in 2018
- After 2018, the success rate increases back

All Launch Site Names

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

- Using DISTINCT("Launch_Site")
- DISTINCT will filter the unique values in the column
- There are 4 different launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Using WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
- WHERE clause is used to set the condition
- LIMIT is to restrict the number of result showed

Total Payload Mass

SUM("PAYLOAD_MASS__KG_")
619967

- Using SUM("PAYLOAD_MASS__KG_")
- SUM will add up all the values in the column

Average Payload Mass by F9 v1.1

AVG_PAYLOAD_MASS_KG
2928.4

- Using AVG(“PAYLOAD_MASS_KG”) and WHERE "Booster_Version" == "F9 v1.1"
- AVG will find the mean of the column
- WHERE clause is used to set the condition to match the booster version

First Successful Ground Landing Date

MIN("Date")
2015-12-22

- Using MIN("Date") and WHERE "Landing_Outcome" LIKE "Success%"
- MIN is used to find the earliest date
- WHERE is used to set the condition
- LIKE will find all the values begin with “Success”

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Using WHERE "Landing_Outcome" == "Success (drone ship)" AND ("PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000)
- WHERE clause has 2 conditions:

1. Find the corresponding landing outcome
2. Set the range for payload mass

Total Number of Successful and Failure Mission Outcomes

Outcome_Cat	Total
Failure	10
Other	30
Success	61

- Using CASE clause to set 3 cases:

1. When outcome contains “Success” -> Success
2. When outcome contains “Failure” -> Failure
3. Other values in outcome -> Other

- Using GROUP BY to count the number and group according to the cases

Boosters Carried Maximum Payload

Booster_Version
F9 B5 B1048.4

- Using ORDER BY “PAYLOAD_MASS_KG” DESC LIMIT 1
- ORDER BY will arrange the value in the column in ascending / descending order
- LIMIT will restrict the number of result displayed

2015 Launch Records

Month	Landing_Outcome	Booster_Version	"Lauch_Site"
01	Failure (drone ship)	F9 v1.1 B1012	Lauch_Site
04	Failure (drone ship)	F9 v1.1 B1015	Lauch_Site

- Using substr(Date,6,2) to get the Month and substr(Date,0,5) = "2015“ to get the Year
- Using WHERE clause to set 2 conditions:

1. Filtered Landing_Outcomes to desired value
2. Set the year to be 2015

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

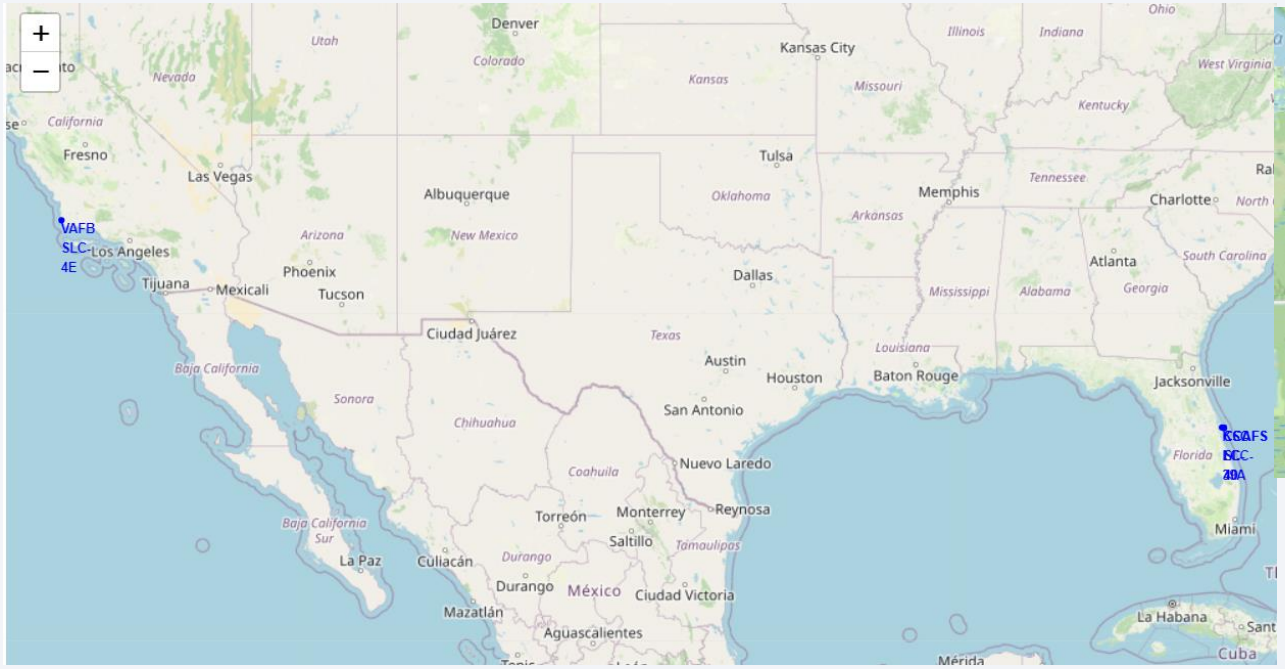
Landing_Outcome	COUNT("Landing_Outcome")
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

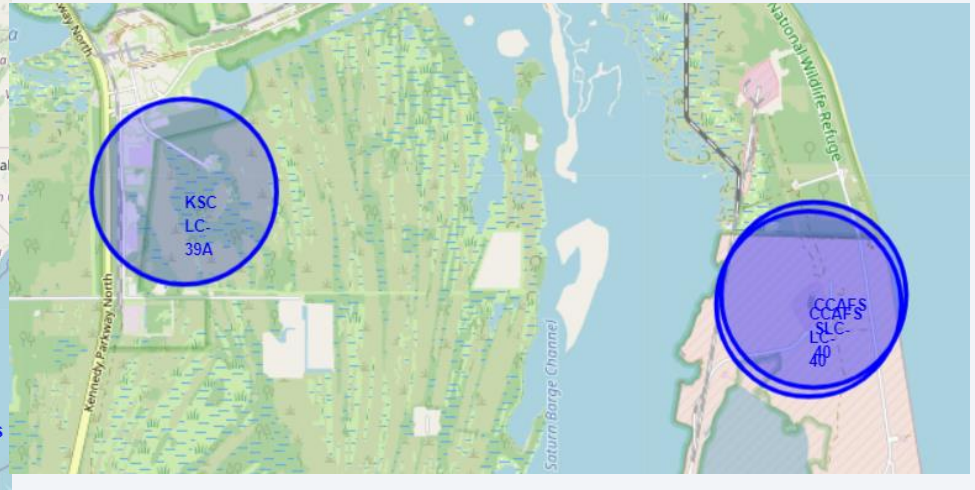
- Using COUNT, WHERE, GROUP BY and ORDER BY clause
- COUNT the Landing Outcome according the the condition set by WHERE clause and GROUP BY clause
- Order the result in descending order
- No attempt accounted the most for the landing outcome

Section 3

Launch Sites Proximities Analysis

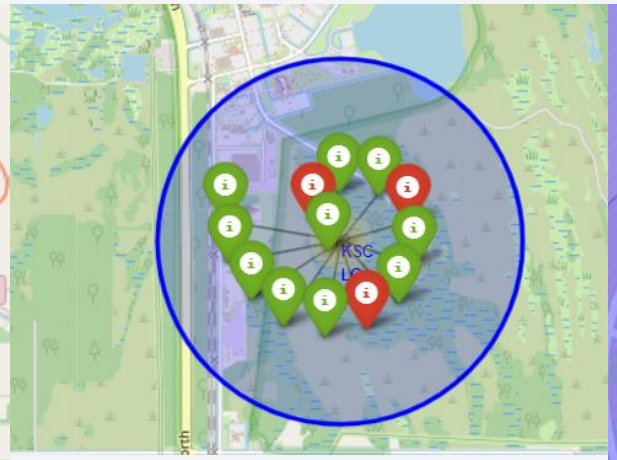
Launch Sites on Map

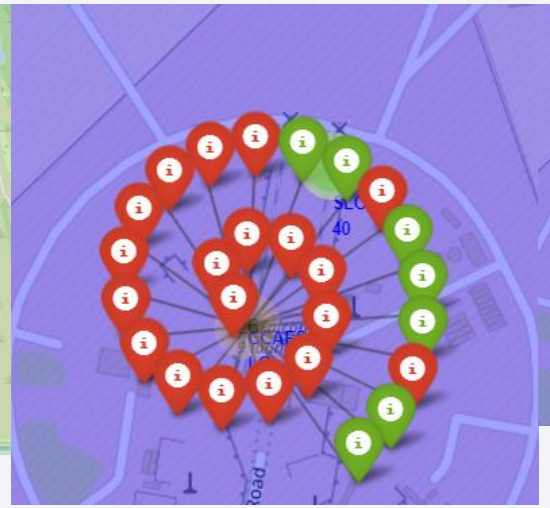


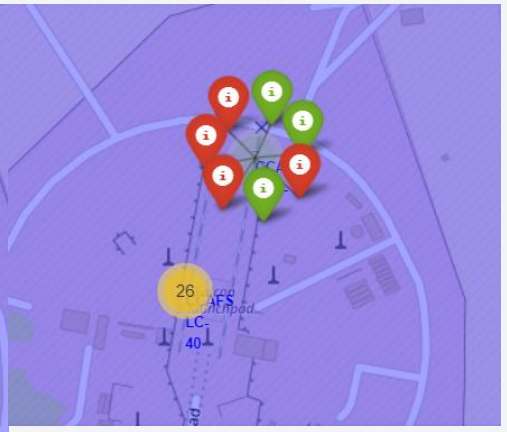


- 1 launch site is at the West, the other 3 is on the East
- All the launch sites are near to the coastline

Success / Fail Launches of Each Site

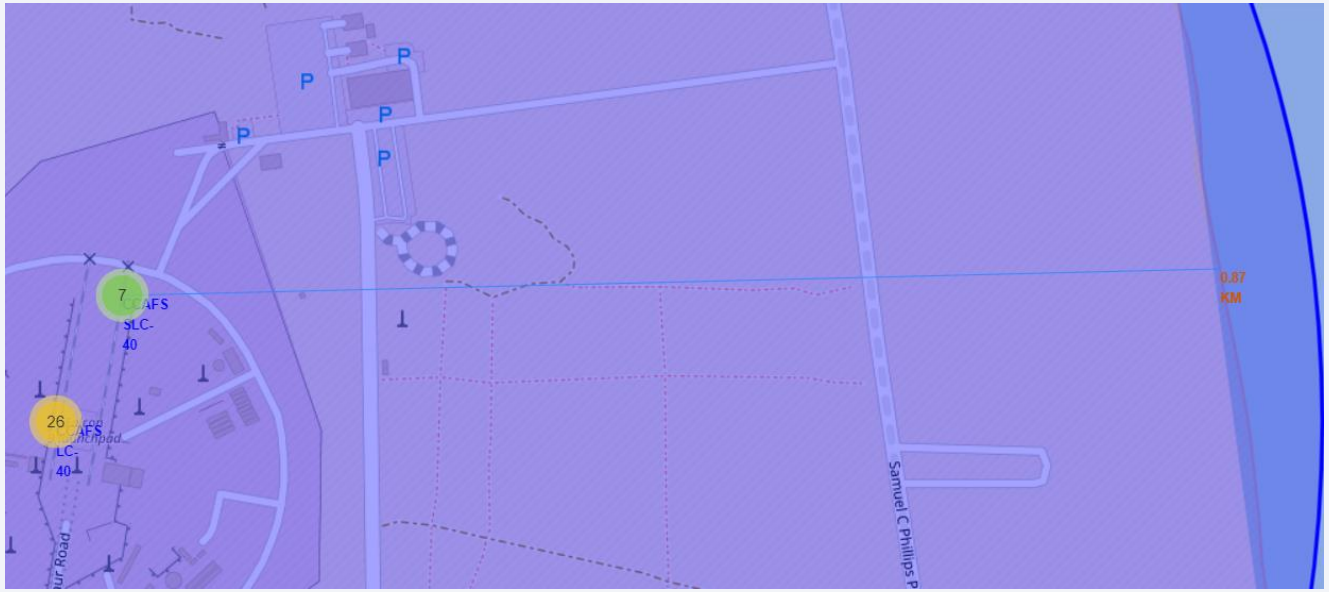




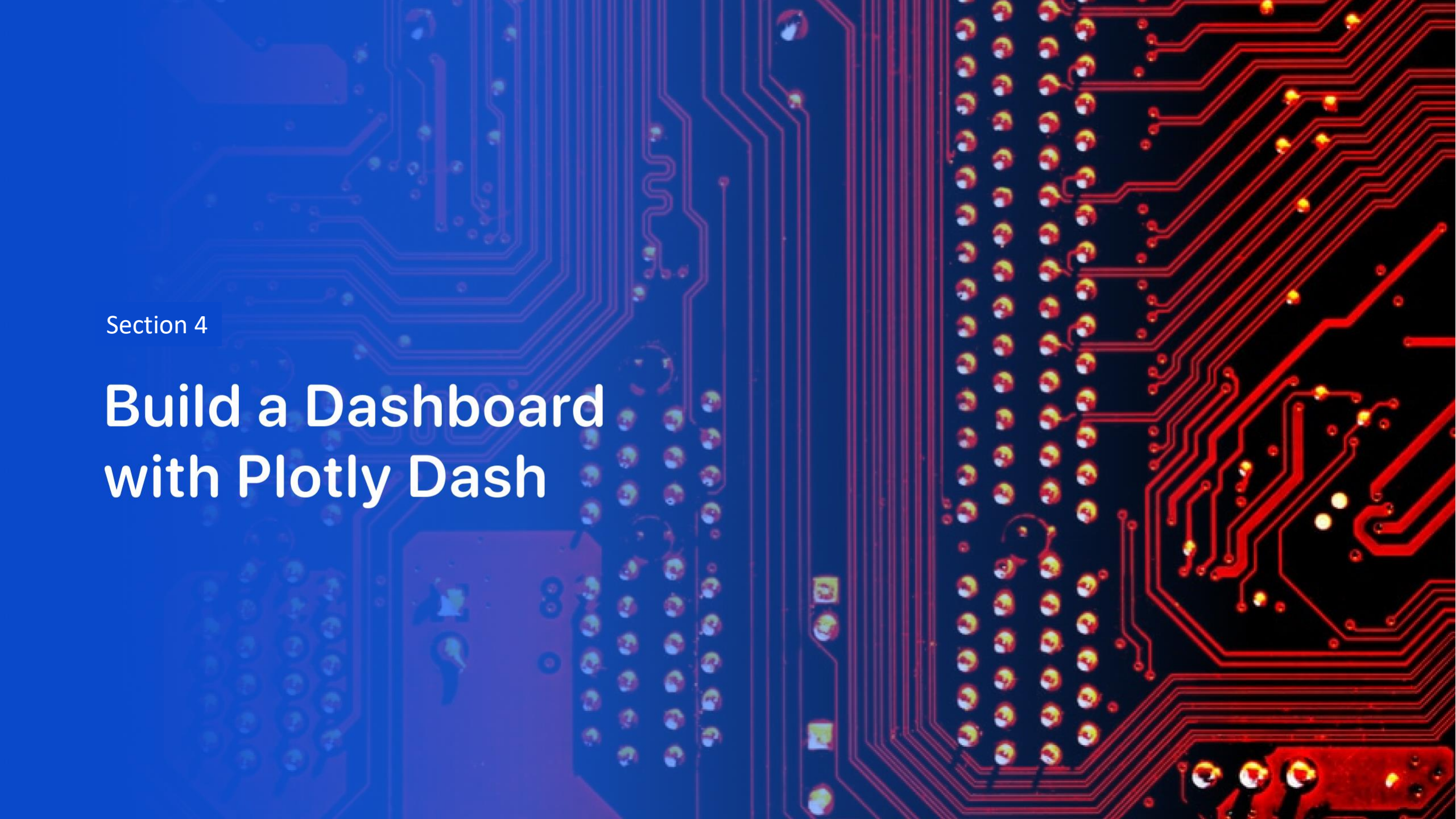


- KSC LC-39A has the highest success rate
- CCAFS LC-40 has the most launches
- CCAFS SLC-40 has the least launches

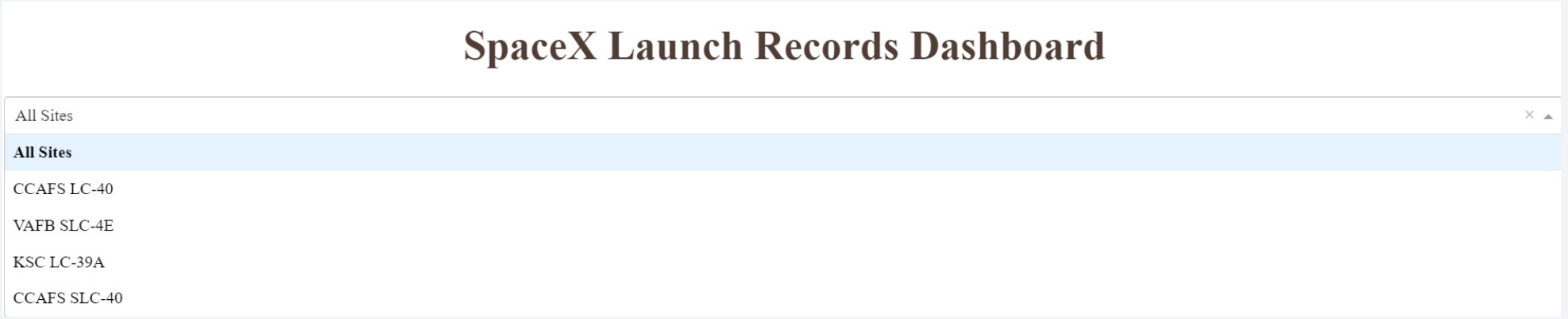
Distance between Launch Site and Coastline



- A line is drawn between the launch site and coastline point
- The distance between them is 0.87km

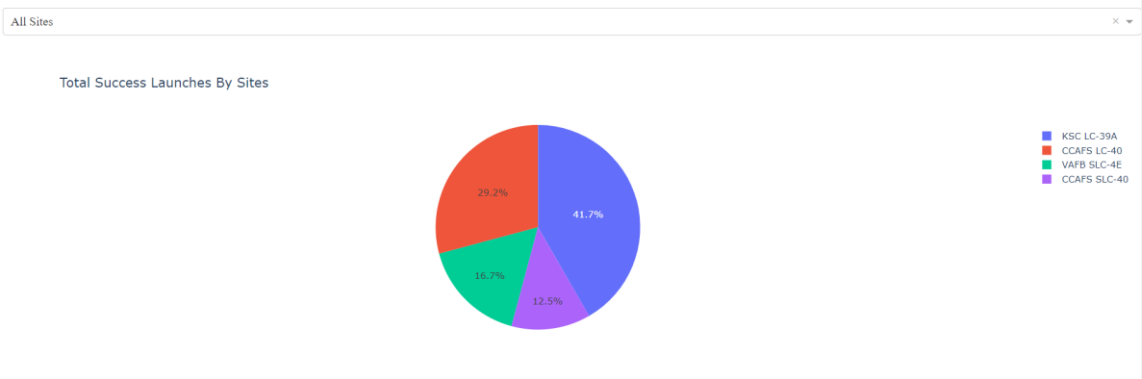

Section 4
Build a Dashboard with Plotly Dash

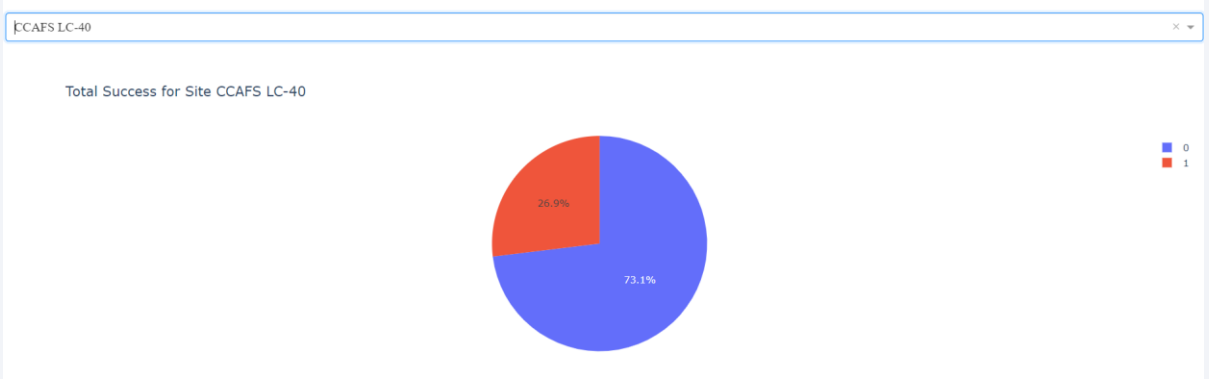
Launch Site Selection



- User can select the one specific launch site or all launch sites

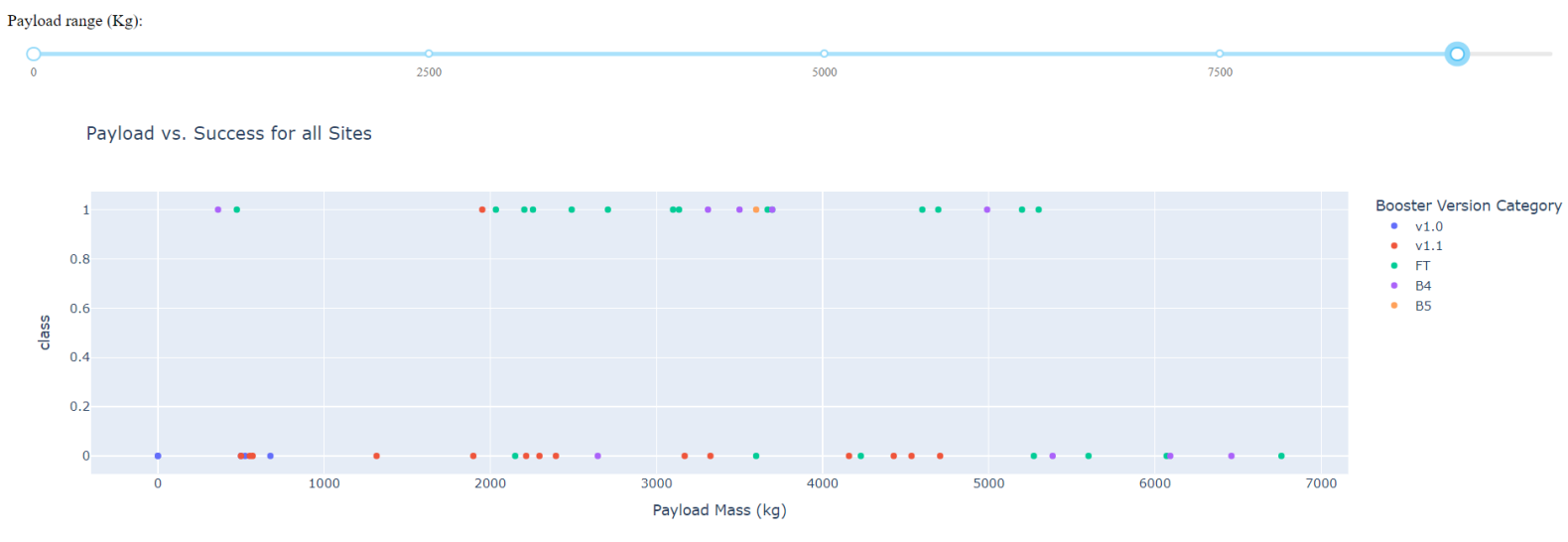
Success Rate Pie Chart





- Left side shows the success rate distribution of all launch sites
- Right side shows the success rate of one specific launch site, where the red colour denotes success

Scatter Chart of Payload Mass vs Launch Outcome

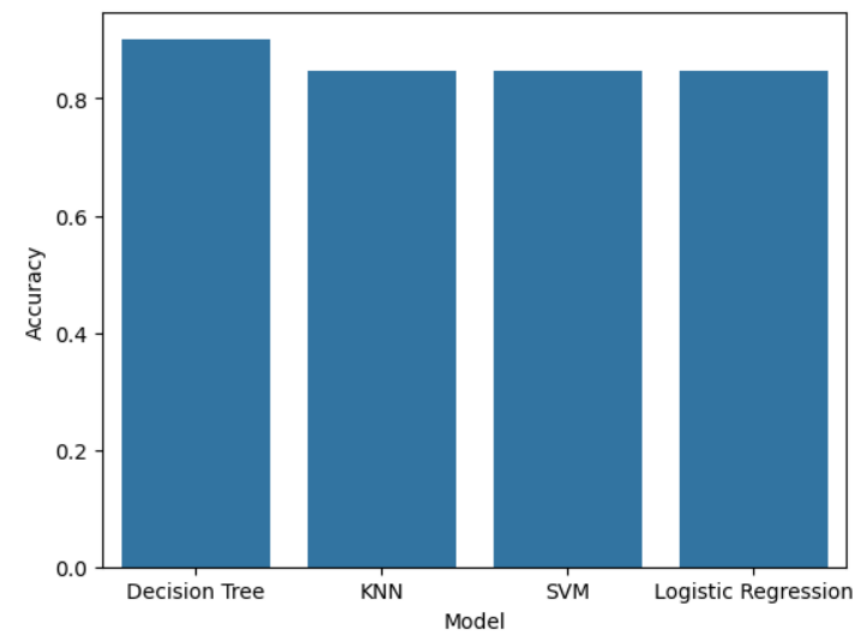


- FT booster has the highest success rate
- V1.1 booster has the lowest success rate
- The payload mass range is set between the minimum and maximum value

Section 5

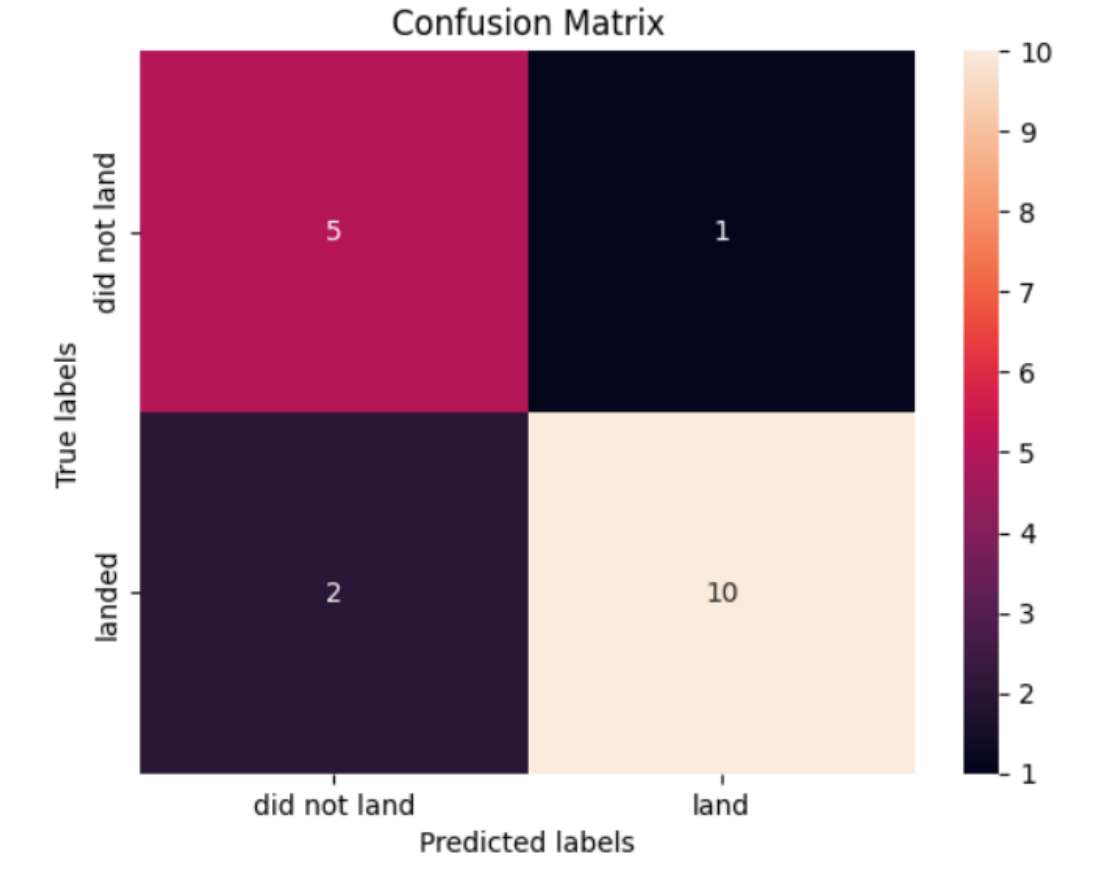
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree ahs the highest accuracy
- KNN, SVM and Logistic Regression has not much difference in their accuracy

Confusion Matrix



- Out of 6 samples in “did not land”, 5 are predicted correctly and 1 is predicted wrong (False Positive)
- Out of 12 samples in “landed”, 10 are predicted correctly, and 2 is predicted wrongly (False Negative)

Conclusions

- There are 4 different launch sites
- The yearly trend of success rate shows an increasing trend
- ES-L1, GEO, HEO and SSO have the highest success rate
- Launch site KSC LC-39A has the highest success rate
- FT booster has the highest success rate
- The best prediction model is decision tree

Appendix

- Github Repository

Thank you!