IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

I. Methodologies:

- Encode categorical features using One Hot Encoder
- Predict a landing’s success using basic machine learning algorithms: Logistic Regression, Support Vector Machines, K Nearest Neighbors and Multinominal Naïve Bayes

II. Result:

- The highest test accuracy is 94.44% and belongs to the Decision Tree model
- The major problem of all models are the number of false positives

Introduction

- Space travel is becoming commercially available
- Landing successful ➔ first stage reused ➔ travel's cost considerably saved
- Our problem: determining whether a landing is successful or not based on multiple factors:
    - Payload's mass
    - Orbit type
    - Launch site
    - Number of flights
    - Having grid fins or not
    - Reused counting
    - Block number
    - Landing pad's code

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

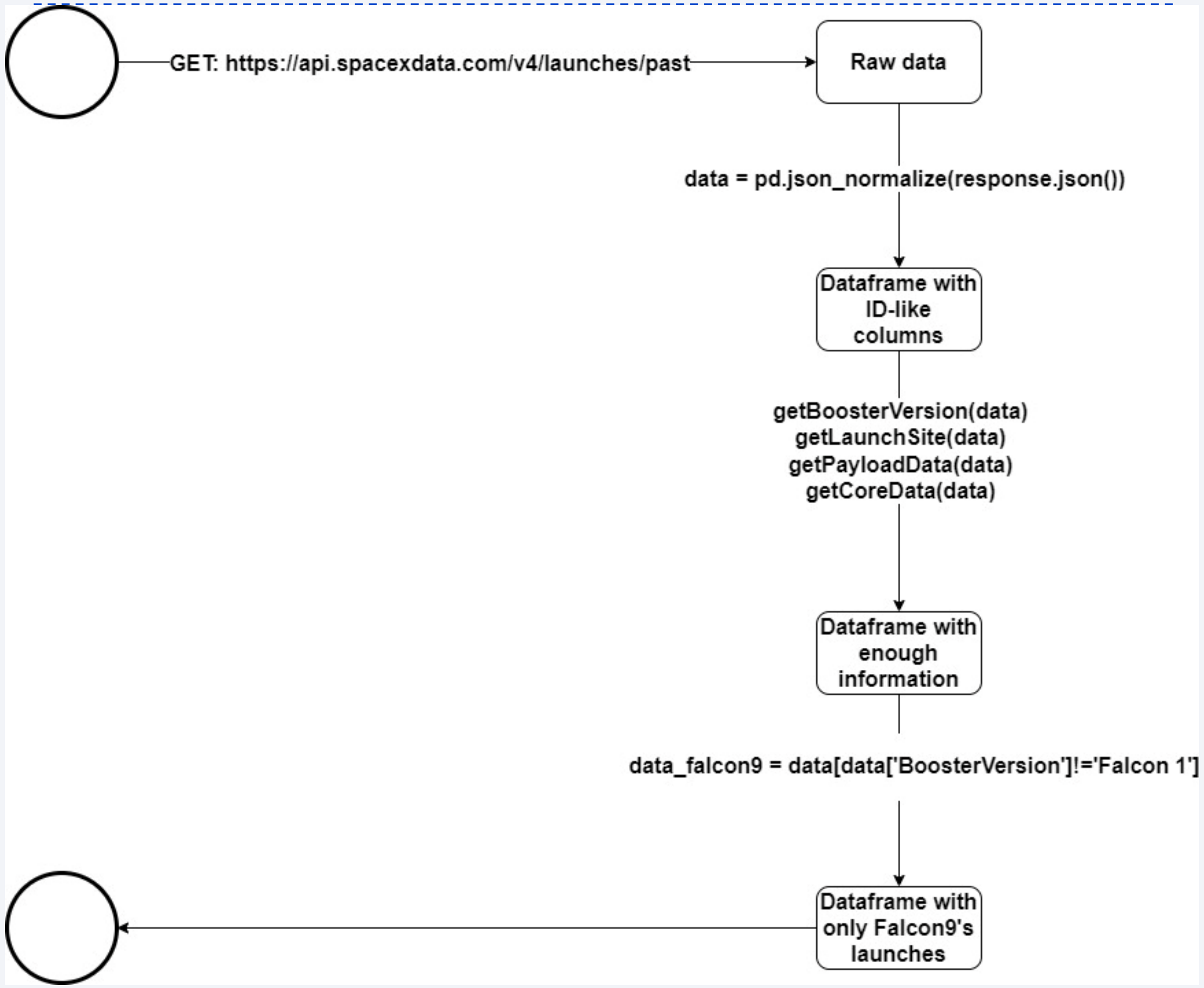
Data Collection

- Dataset used: SpaceX Launch Data
- Data is collected by calling SpaceX's API, i.e. making a GET request to https://api.spacexdata.com/v4/[endpoint]
- [endpoint] is:
  - /launches/past: get comprehensive past launch data and is the need to get data from the below endpoints
  - /rockets/: get booster's name
  - /launchpads/: get launch site's name, latitude and longtitude of each launch
  - /payloads/: payload's mass and orbit of each launch
  - /cores/: landing's outcome, number of flights, block number, etc.

Data Collection – SpaceX API

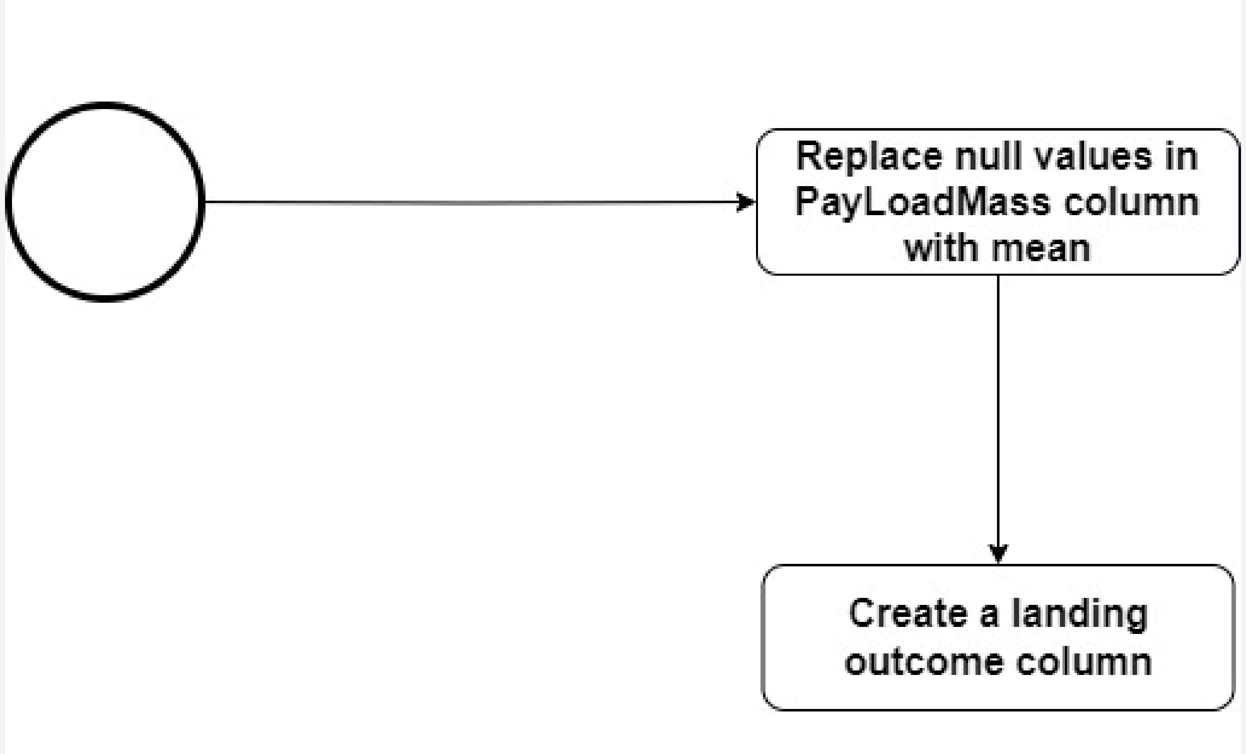
- GET request ➔Raw data ➔ Some converts ➔ Final dataframe

- Reference [this notebook]



Data Wrangling

- Calculating the number of null values in each column
- Converting null value to the mean of that column
- Reference this notebook



EDA with Data Visualization

- Categorical plots: to see the relationship between FlightNumber, PayloadMass, LaunchSite and the launch outcome
- Bar chart: to see the relationship between success rate and orbit type
- Scatter plot: to see the relationship between FlightNumber, PayloadMass, Orbit and the launch outcome
- Line plot: to see the success rate from 2010 to 2020

- Reference this notebook

EDA with SQL

- Display unique launch sites
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List booster versions which have success in drone ship and have payload mass in range(4K, 6K)
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List months, failure landing_outcomes in drone ship ,booster versions, launch_site for year 2015
- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order
- Reference this notebook

Build an Interactive Map with Folium

- Circle: highlight an area with specific coordinates, e.g. NASA Johnson Space Center
- Marker: add text label to a Circle object or a PolyLine object
- Circle group: add a circle polygon surrounding an area
- Marker cluster: group markers with the same coordinates
- PolyLine: draw a line between two destinations

- Reference [this notebook]

Build a Dashboard with Plotly Dash

- Dropdown list: offer various launch site options for users to select
- Pie chart:
    - If "All Sites" is selected, represent the total number of successful landings in all launch sites
    - If a specific launch site is selected, represent the successful/failed landing ratio and the number of them
- Slider: to choose a payload range to plot a scatter plot of PayloadMass vs the landing's outcome
- Scatter plot: to illustrate the relationship between PayloadMass and the landing's outcome
- Reference [this Python script]

Predictive Analysis (Classification)

- For each model, set a value range for its hyperparameters ➔ find the set of hyperparameters that yield the best accuracy on training set
- Grid Search Cross Validation is employed:
  - Set cv = 10, param_grid = {param1: [value1.1, value1.2],
param2: [value2.1, value2.2],
…}
  - For each hyperparameter, dataset is divided into 10 subsets. The training phase will take 10 times, each time one subset acts as the testing set
  - Calculate the average test score of 10 times

- Reference this notebook

Results

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

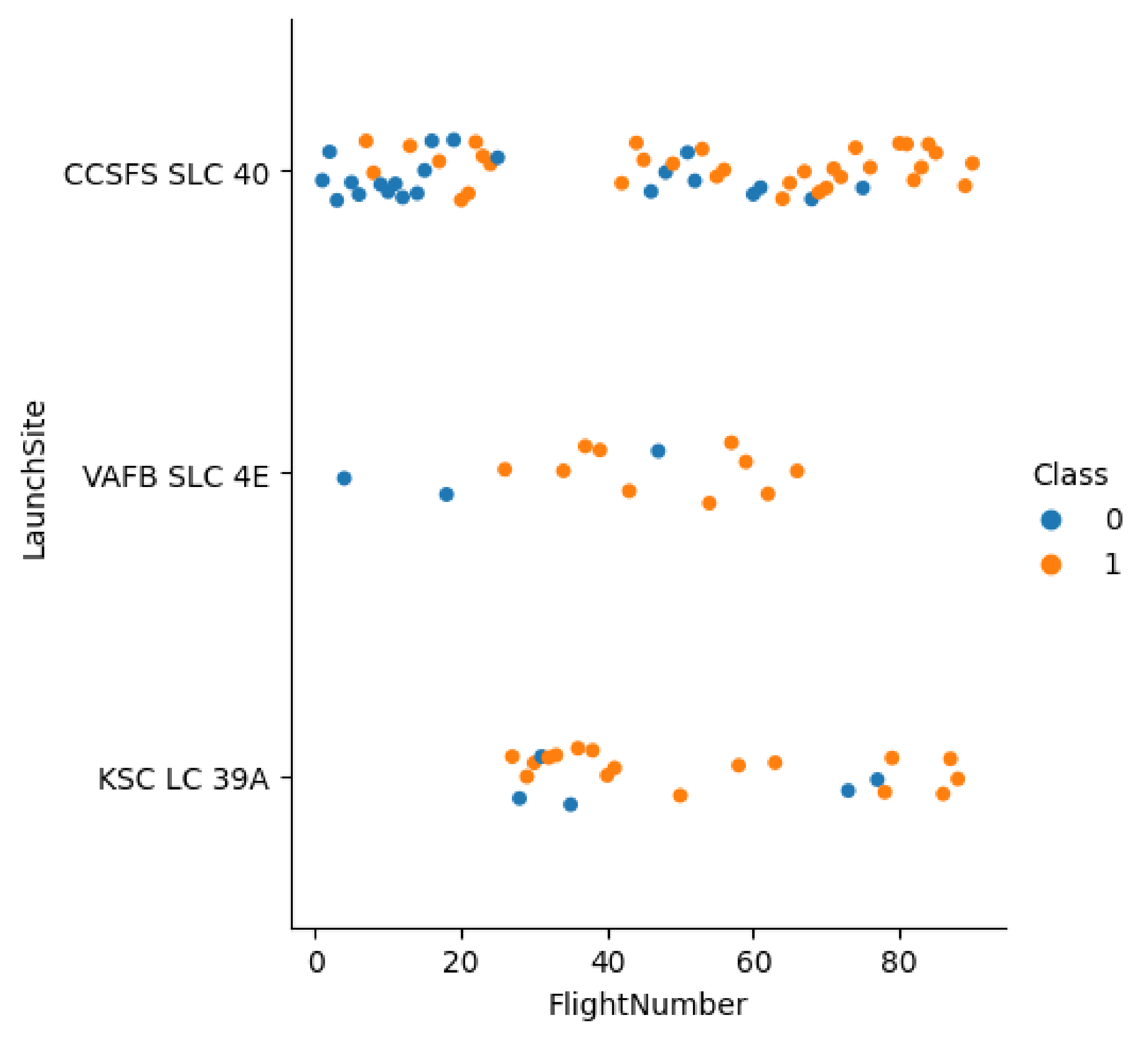
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

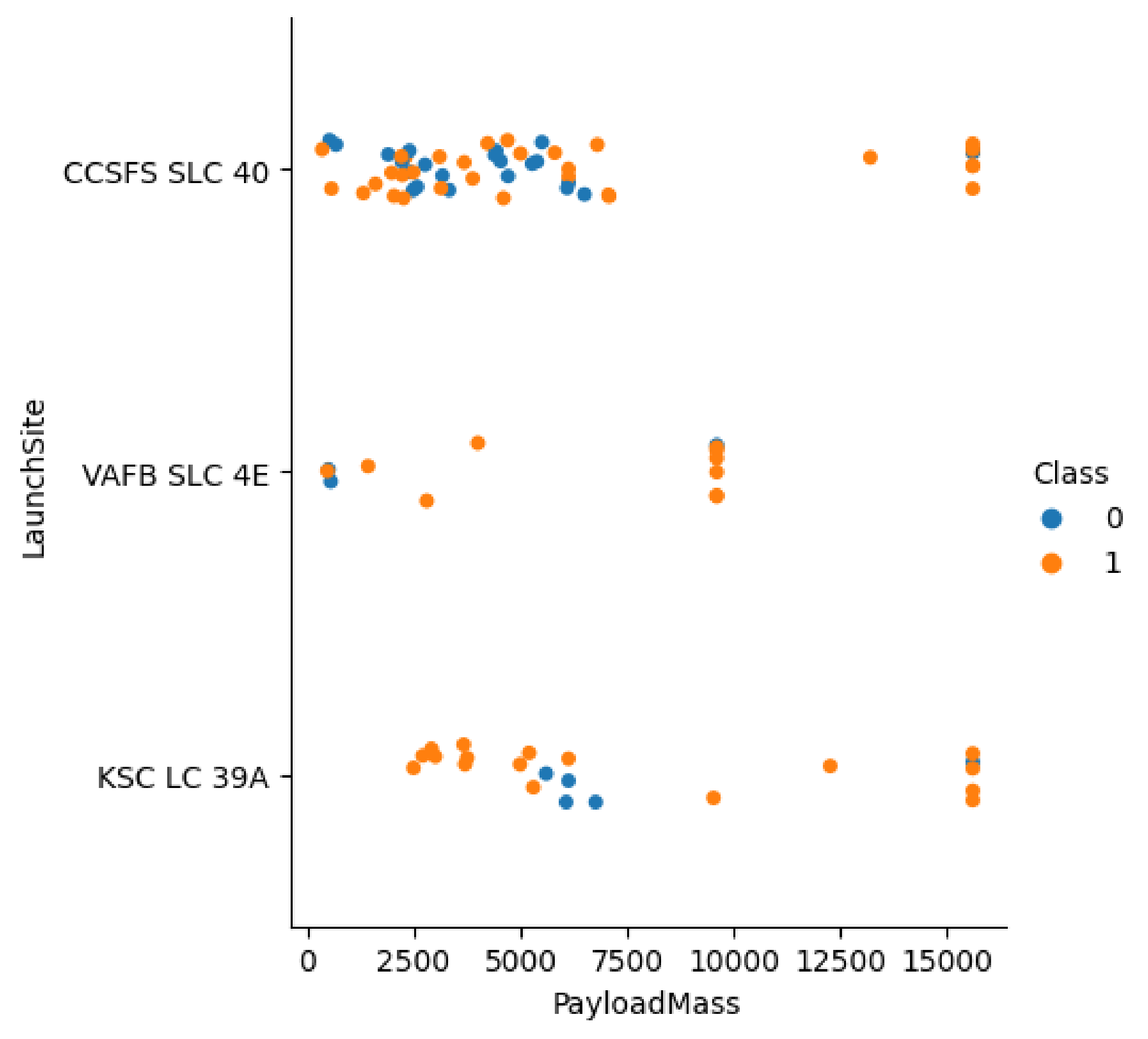
- In the latter flights, the team responsible for a missile program has more experience and improve many aspects of the rocket

➔ success rate higher than that of the former ones



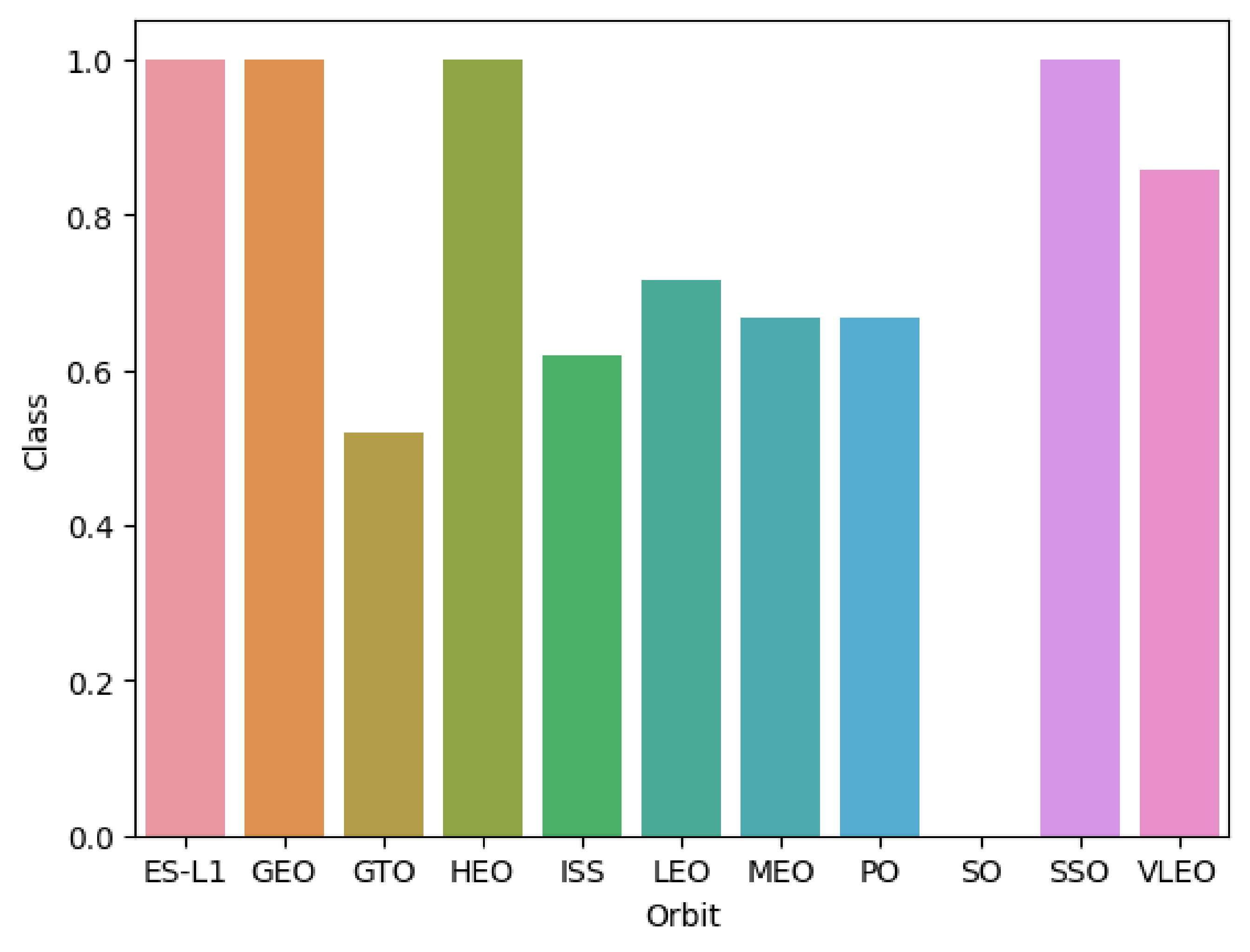
Payload vs. Launch Site

- Higher payload ➔ extensive testing and scrutiny ➔ risks minimized ➔ higher success rate
- VAFB SLC-4E is often employed for Earth observation satellites and other scientific missions, which tend to have payloads up to 10,000 kg



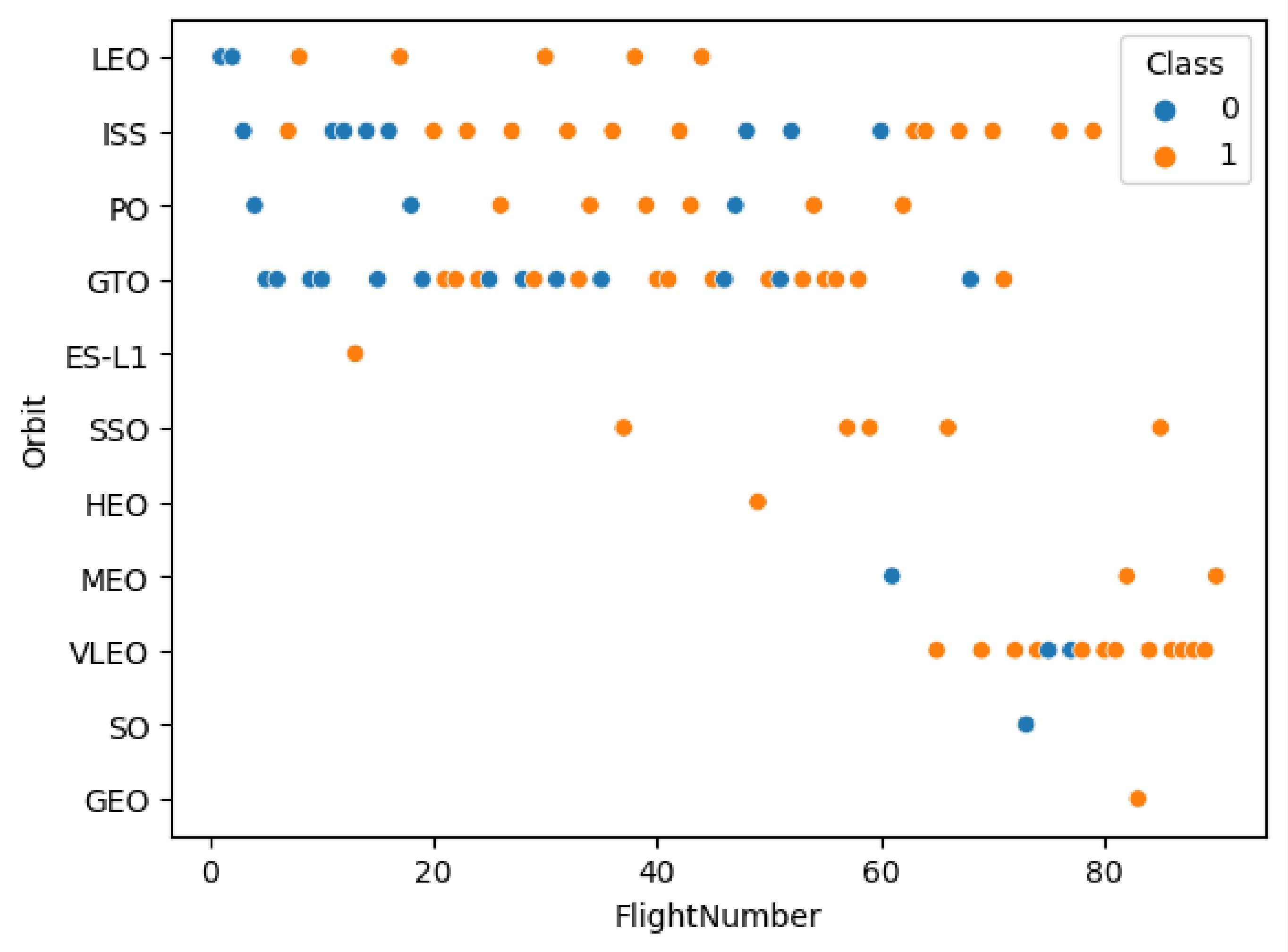
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have approx. 100% success rate
- Lowest success rate, which is approx. 0%, belongs to SO orbit



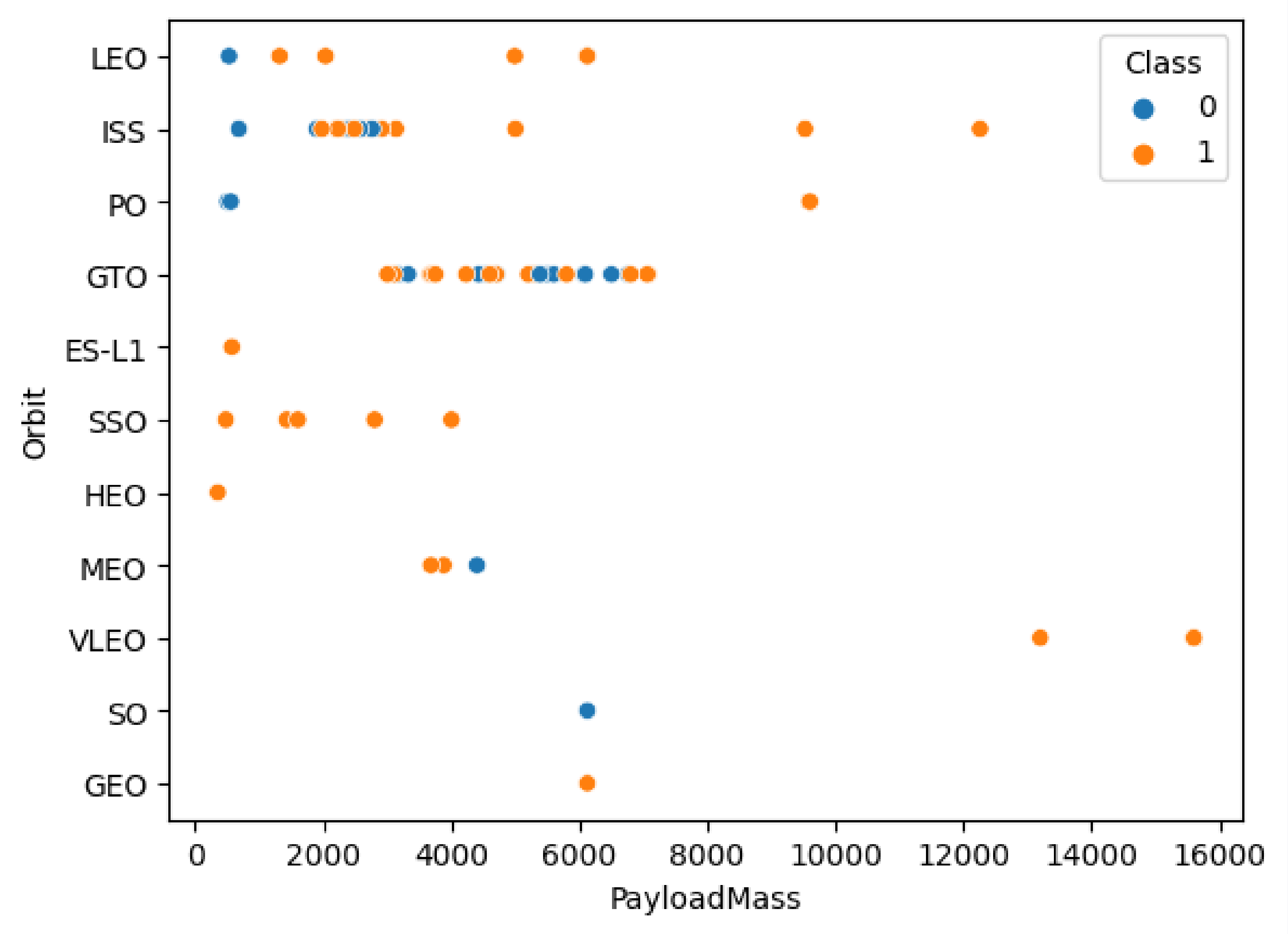
Flight Number vs. Orbit Type

- First launches tend to expose to failure ➔ more experience, more improvement on the rocket's design and payload ➔ following launches have more chance to succeed
- Launches with flight number > 80 always succeed
- SSO, HEO and ES-L1 orbits have 100% successful launch rate, but with very little total number of launches
- VLEO is only employed for flight with order > 60



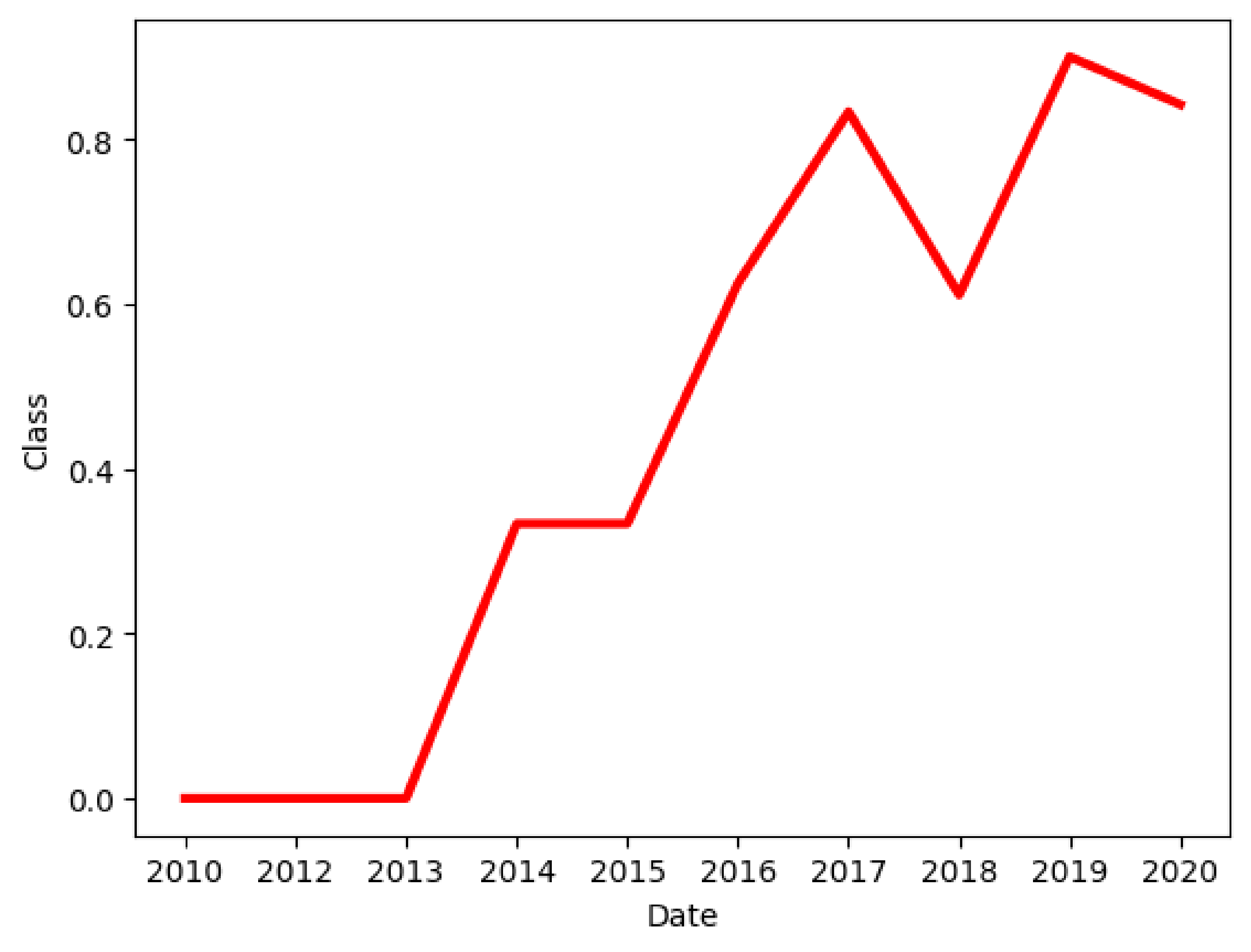
Payload vs. Orbit Type

- With heavy payloads the successful landing rate are more for PO, LEO and ISS orbits



Launch Success Yearly Trend

- SpaceX's rockets have been strengthened through years ➔ increasingly high success rate
- Notable decreases occur between 2017 and 2018, 2019 and 2020



All Launch Site Names

- Select distinct values from Launch_Site column
- There are some rows in which a launch site is not specified

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40
4	None

Launch Site Names Begin with 'CCA'

- Launch site begins with 'CCA' ➔ Launch site has a pattern of 'CCA%'
- 5 records ➔ 5 rows ➔ limit 5

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
1	12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
3	10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
4	03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- 98 rows ➔ 98 different booster versions
- For each booster version, sum all of its launches' payload ➔ sum(PAYLOAD_MASS__KG_)

	Booster_Version	sum(PAYLOAD_MASS__KG_)
0	None	NaN
1	F9 B4 B1039.2	2647.0
2	F9 B4 B1040.2	5384.0
3	F9 B4 B1041.2	9600.0
4	F9 B4 B1043.2	6460.0
...	...	...
93	F9 v1.1 B1014	4159.0
94	F9 v1.1 B1015	1898.0
95	F9 v1.1 B1016	4707.0
96	F9 v1.1 B1017	553.0
97	F9 v1.1 B1018	1952.0

98 rows × 2 columns

Average Payload Mass by F9 v1.1

- Sum all launches' payload carried by F9 v1.1 / number of launches boosted by F9 v1.1

➔ Average Payload Mass by F9 v1.1

	avg(PAYLOAD_MASS__KG_)
0	2928.4

First Successful Ground Landing Date

- The *Date* column is already in an ascending order
- Select rowshaving Landing_Outcome = “Success (ground pad)”
- Limit to 1 row

	Date	Landing_Outcome
0	22/12/2015	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Select rows satisfying two conditions:
  - Landing_Outcome = “Success (drone ship)”
  - Payload_Mass__KG_ between 4000 and 6000

	Booster_Version	Landing_Outcome
0	F9 FT B1022	Success (drone ship)
1	F9 FT B1026	Success (drone ship)
2	F9 FT B1021.2	Success (drone ship)
3	F9 FT B1031.2	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Using aggregate function: *Count(*)*
- Result grouped by *Mission_Outcome* column

	Mission_Outcome	Number
0	None	0
1	Failure (in flight)	1
2	Success	98
3	Success	1
4	Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Select rows satisfying:
  - PAYLOAD_MASS__KG_ = select max(PAYLOAD_MASS__KG_)

	Booster
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

- Select rows satisfying:
  - substr(Date, 7, 4) = ‘2015’ (year 2015)
  - Landing_Outcome = ‘Failure (drone ship)’
- In, there are 2 drone-ship landings which are failed and both took place in CCAFS LC-40

	Month	Booster_Version	Launch_Site
0	10	F9 v1.1 B1012	CCAFS LC-40
1	04	F9 v1.1 B1015	CCAFS LC-40

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

- Select Landing_Outcome having Date between 2010-06-04 and 2017-03-20
- Group up all types of landing outcome and count the number of landings in each type
- Order by the counting in descending order

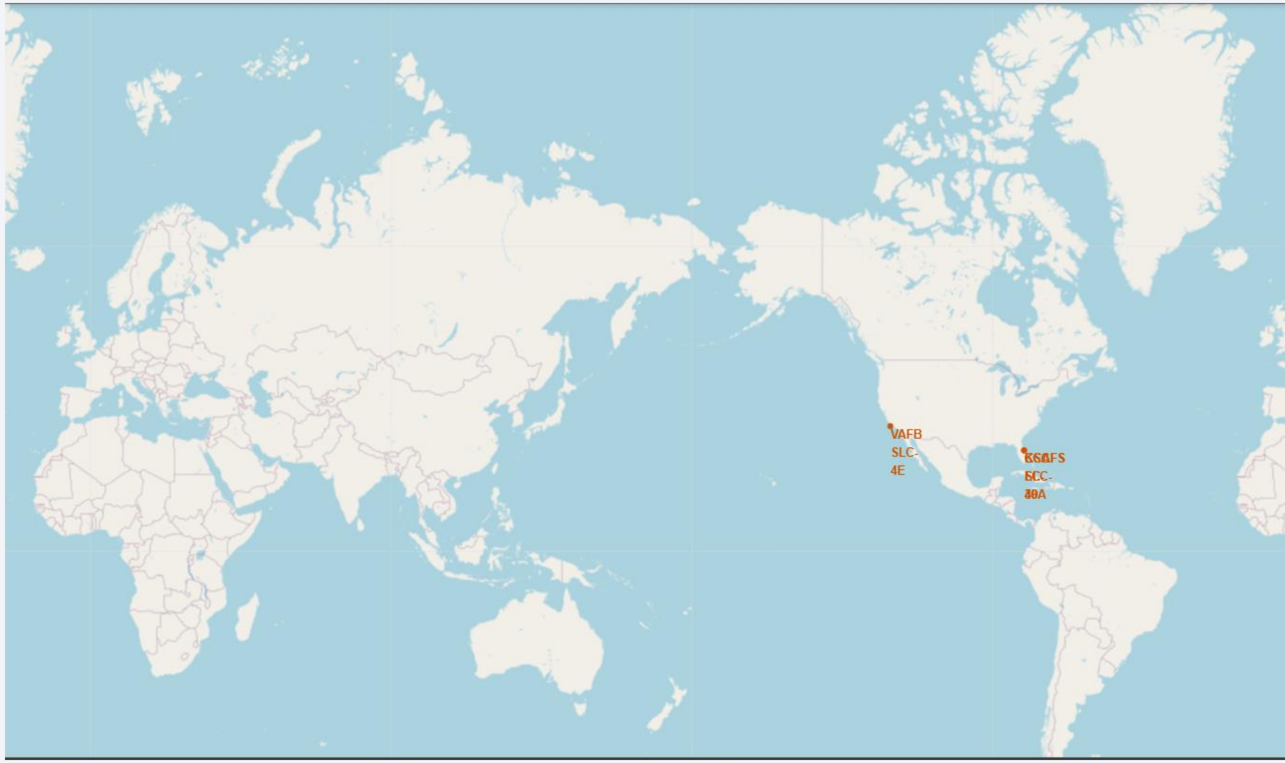
	Landing_Outcome	Counting
0	No attempt	10
1	Success (ground pad)	5
2	Success (drone ship)	5
3	Failure (drone ship)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	2
7	Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

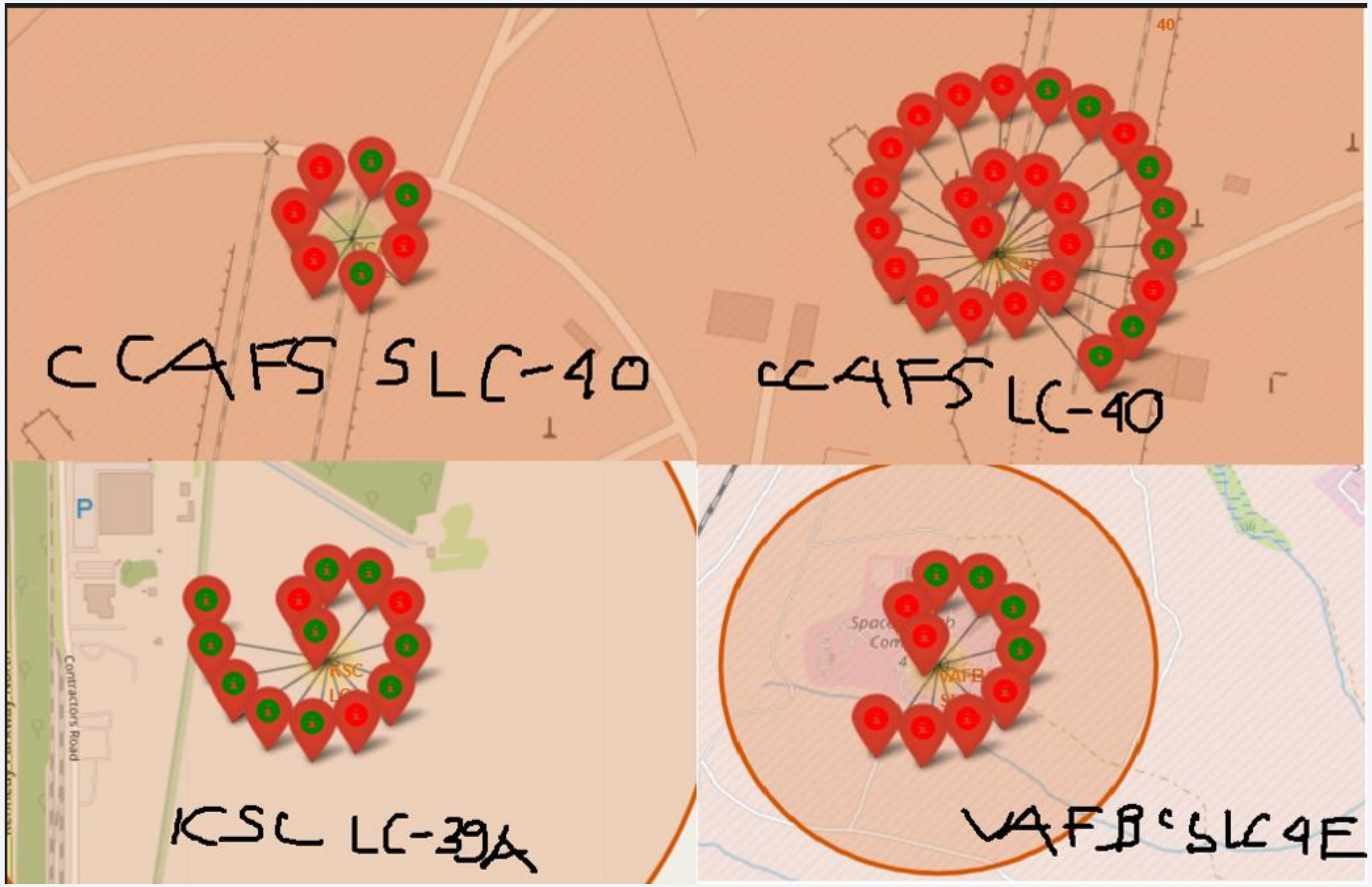
Location of launch sites

- All launch sites are located in close proximity to the nearest coast
- Launch sites in Florida are proximal to each other. This can be explained by these points:
  - Near the equator ➔ bonus velocity during a rocket's launch ➔ improve capability to reach certain orbits
  - Efficient transportation of equipment and personnel between different sites ➔ reduces logistical challenges.



Launch outcomes of each site

- KSC LC-39A has the highest success rate
- CCAFS LC-40 has the highest number of launches but the lowest success rate



CCAFS SLC-40’s distance to certain proximities

- *Samuel C Phillips Parkway* highway provides crucial transportation links for the movement of heavy and oversized equipment and components required for space missions ➔ in close distance (590 meters)
- Coast provides unrestricted airspace and launch safety ➔ in close distance (860 meters)
- While railroad may supplement rocket components and other facilities, for safety it should be relatively far from the launch site ➔ 28 kilometers from *Titan III Railroad*

Section 4

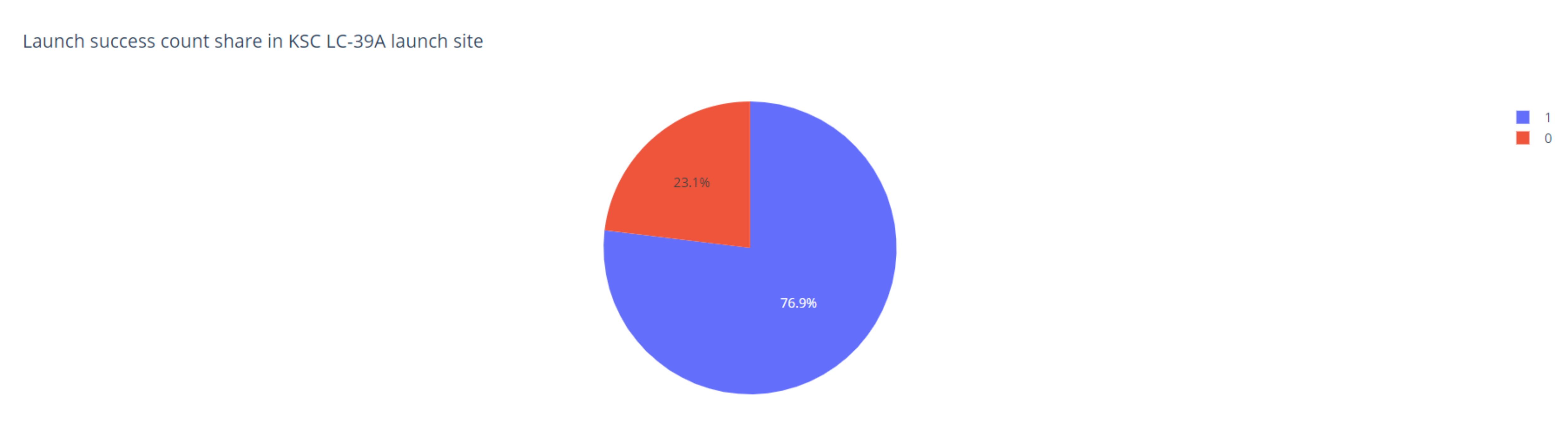
Build a Dashboard with Plotly Dash

Launch success count share between sites

- Florida accounts for 83.4% number of successes, while California makes up 16.7%
- % of flight success in KSC LC-39A = % in CCAFS **LC**-40 + % in CCAFS **SLC**-40
- % of flight success in CCAFS **LC**-40 = % in VAFB SLC-4E + % in CCAFS **SLC**-40

Launch site with highest success rate

- KSC LC-39 A has highest launch success rate, with 76.9% of landings are successful

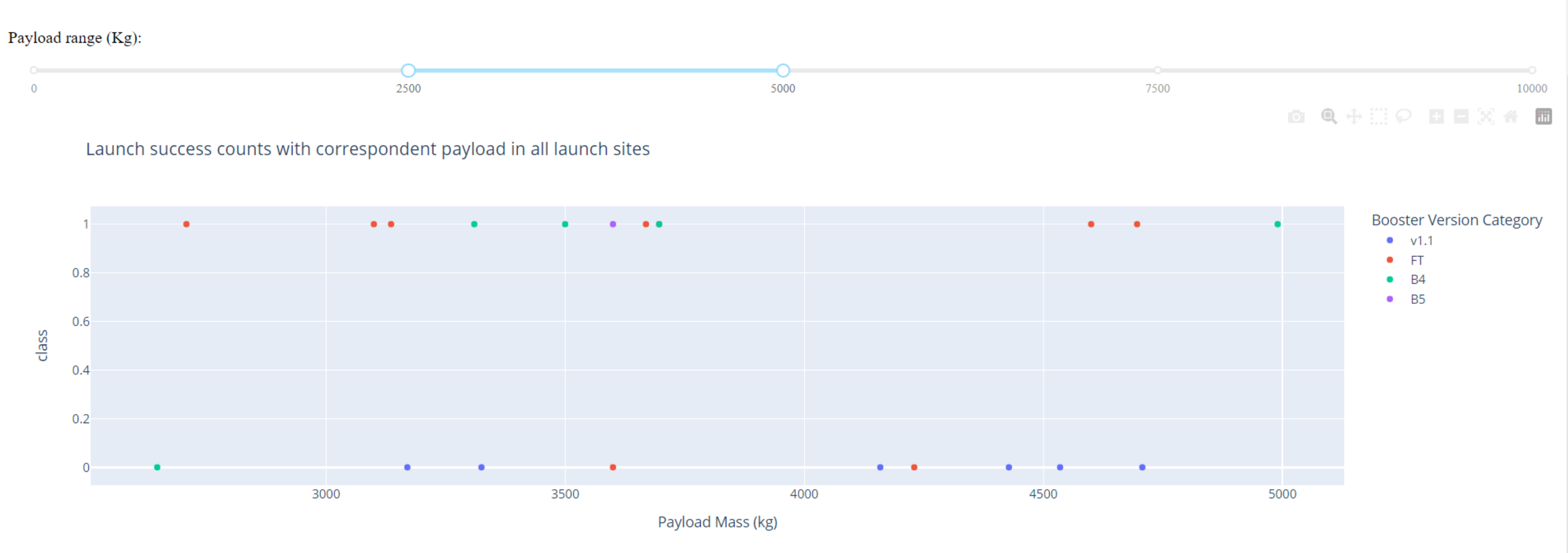


Booster with the highest success rate

- Set the payload range to be 0 to 10000.
- FT has the highest success rate, with 68.42% of flights landed successfully

Booster with the highest success rate

- Payload range from 2500 to 5000 has the highest success rate, with 55% of flights landed successfully

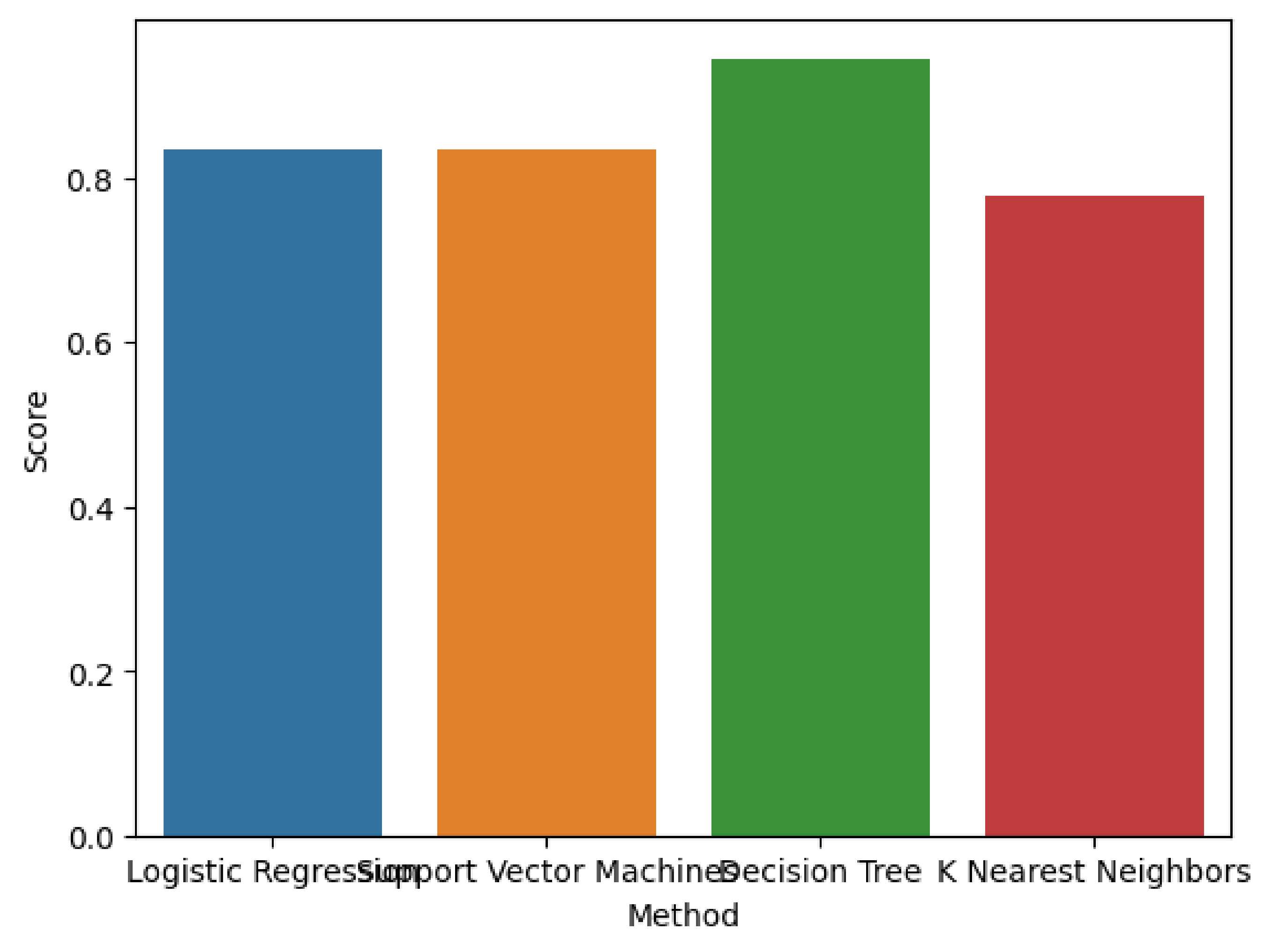


Section 5

Predictive Analysis (Classification)

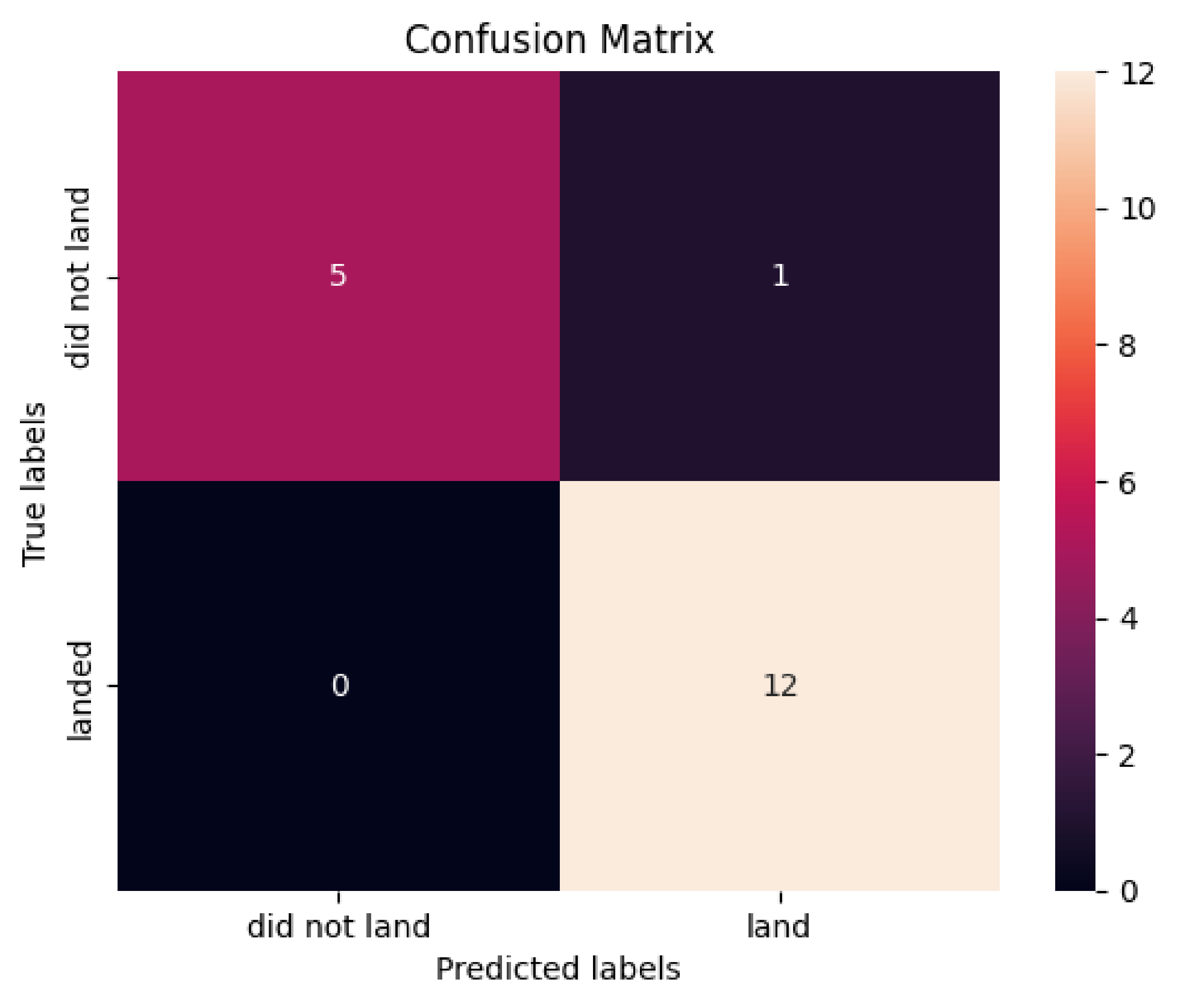
Classification Accuracy

- Decision Tree achieves the highest accuracy on test set, with 94.44% of samples are accurately classified
- K-Nearest-Neighbors, in which k=3, achieves the lowest accuracy on test set (77.78%)



Decision Tree’s confusion matrix

- Only 1 sample is falsely labelled (ground truth: *did not land*, predicted: *land*)
- All samples belonging to *land* class are truly classified



Conclusions

- A rocket's landing outcome can be predicted using machine learning and visualization techniques instead of doing math and physics

➔ Optimize time and resource allocation for launching spaceship

- The correlation between factors is not well-researched enough

➔ Future development: discover more patterns in data and perform more preprocessing techniques to remove unwanted information from the dataset that can badly affect the machine learning model's performance

Thank you!