IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Zhilin Zhao>
<2024.02>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with SpaceX REST API and Web Scraping
  - Exploratory Data Analysis with SQL and Visualization
  - Data Wrangling
  - Interactive Visual Analytics with Folium and Plotly Dashboard
  - Machine Learning Prediction
- Summary of all results
  - The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.

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. 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.
- Problems you want to find answers
  - Can we predict if the Falcon 9 first stage will land successfully?

Section 1

Methodology

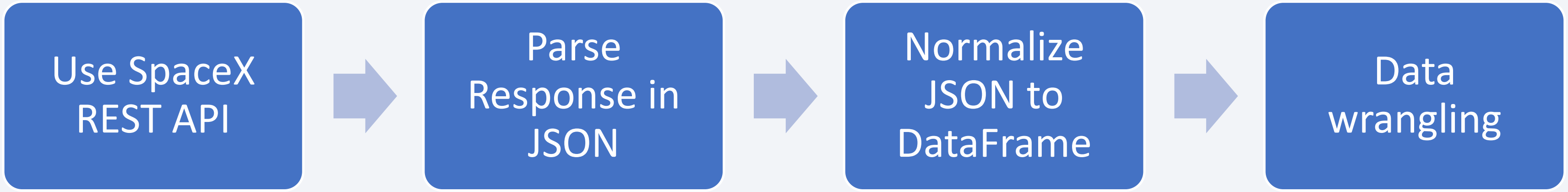
Methodology

Executive Summary

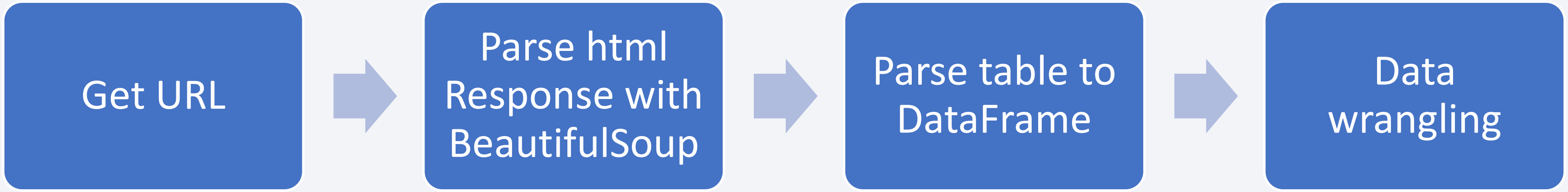
- Data collection methodology:
  - SpaceX RESR API
  - Webscraping with related wiki pages
- Perform data wrangling
  - Filling missing value with `mean`; Converting outcomes into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, KNN, SVM, DT models were built and evaluated for the best model

Data Collection

- SpaceX API.
  -



- WebScraping



Data Collection – SpaceX API

- https://github.com/ringo-alin/IBE-DataScience-Assignments/blob/main/Course%2010/jupyter-labs-spacex-data-collection-api.ipynb

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
# Use json_normalize meethod to convert the json result into a dat
data = pd.json_normalize(response.json())
```

```
# Lets take a subset of our dataframe keeping only the features we
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_nu
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
```

```
# Create a data from launch_dict
data = pd.DataFrame(launch_dict)
```

Data Collection - Scraping

- https://github.com/ringo-alin/IBE-DataScience-Assignments/blob/main/Course%2010/jupyter-labs-webscraping.ipynb

```python
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Fa
```

```python
response = requests.get(static_url).text
```

```python
soup = BeautifulSoup(response)
```

```python
html_tables = soup.find_all('table')
```

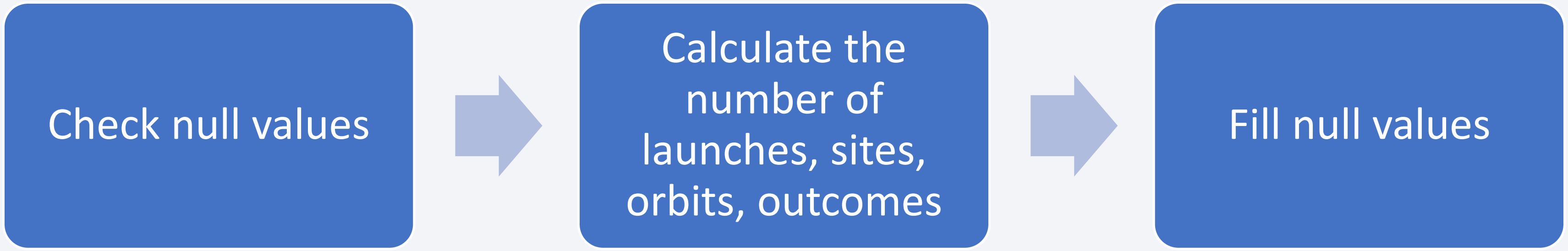
```python
first_launch_table = html_tables[2]
```

```python
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
```

```python
#Extract each table
for table_number,table in enumerate(soup.find_all('table',"wikitabl
    # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is as number correspon
        if rows.th:
            if rows.th.string:
```

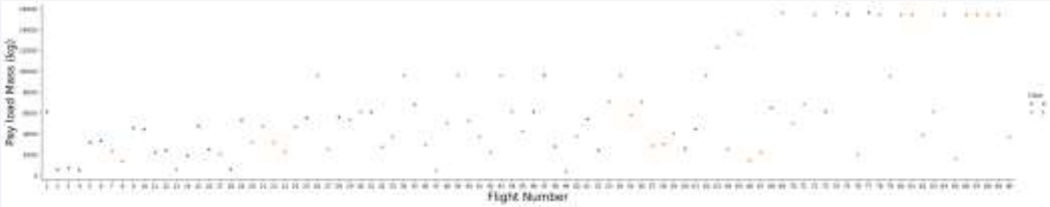
Data Wrangling

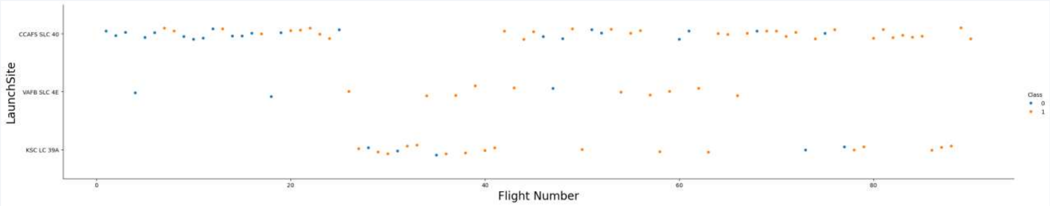
-

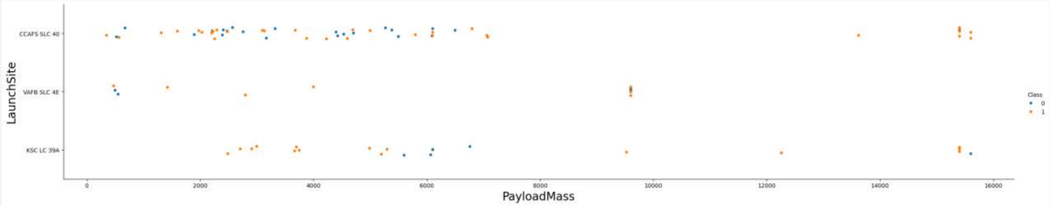


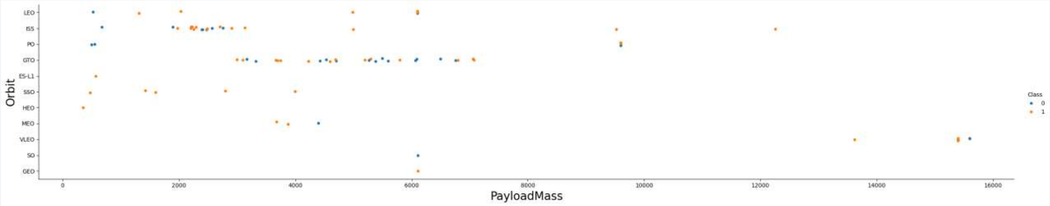
- https://github.com/ringo-alin/IBE-DataScience-Assignments/blob/main/Course%2010/labs-jupyter-spacex-Data%20wrangling.ipynb

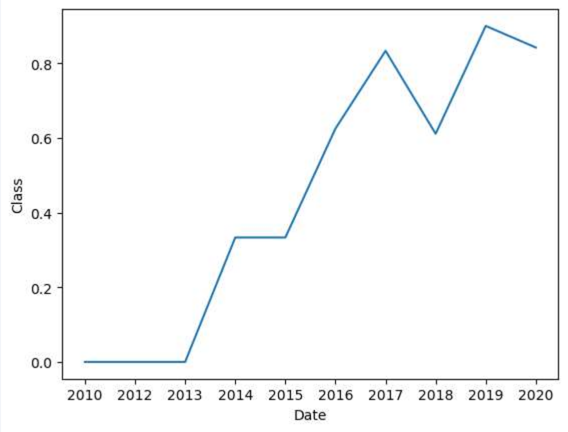
EDA with Data Visualization

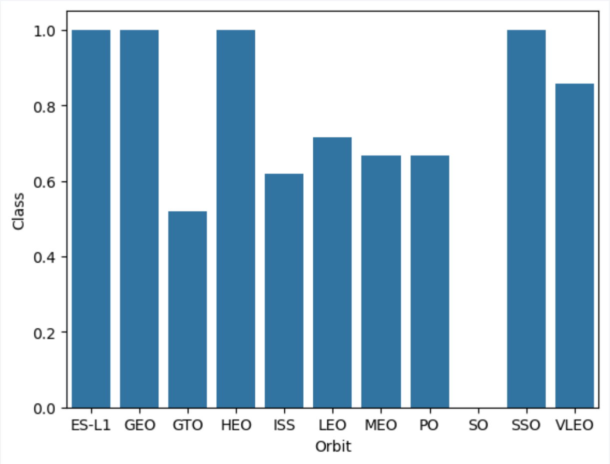


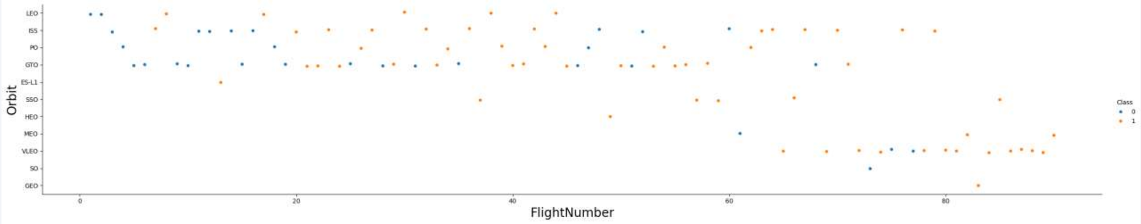












- https://github.com/ringo-alin/IBE-DataScience-Assignments/blob/main/Course%2010/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

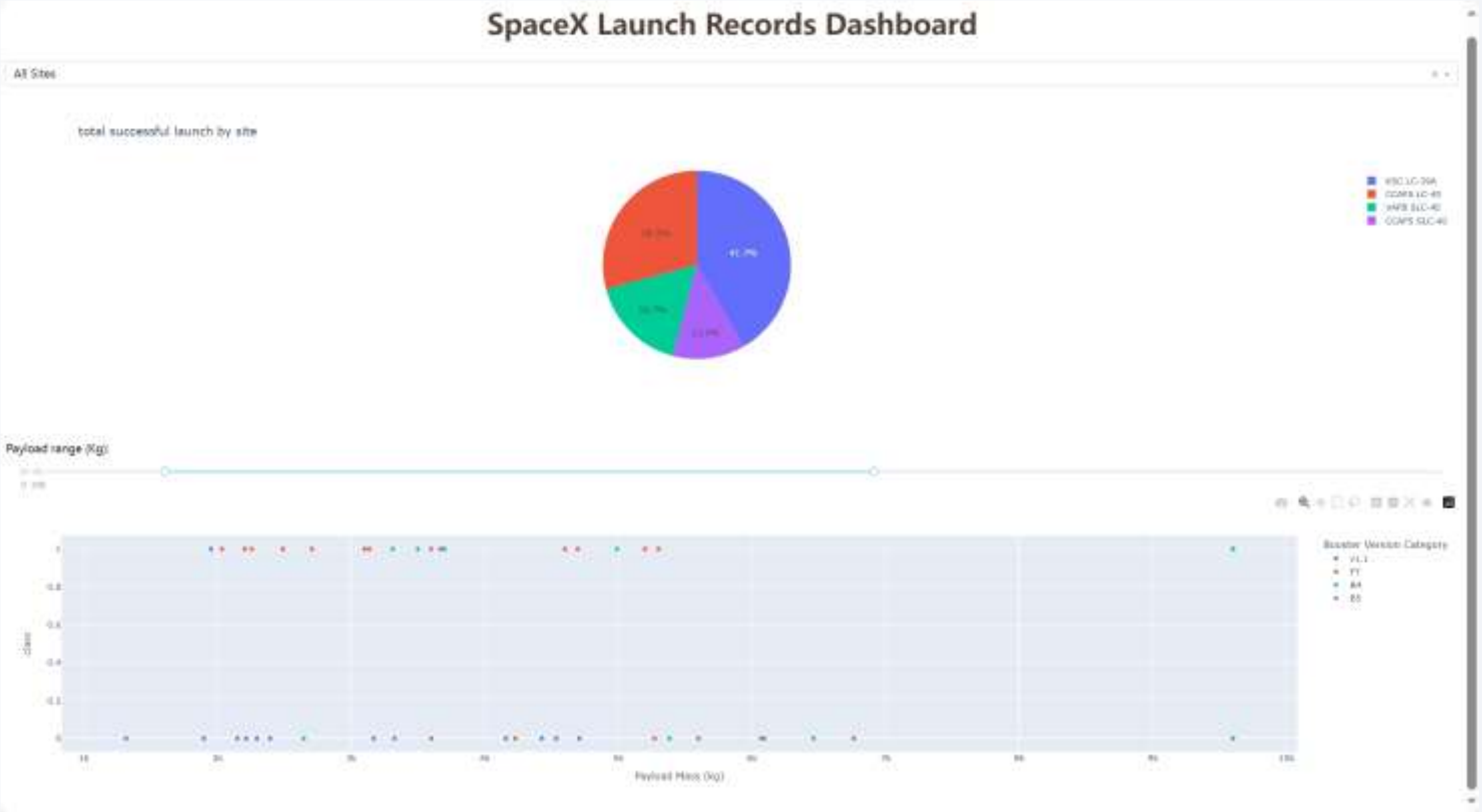
EDA with SQL

- SQL queries
  - select distinct(Launch_Site) from SPACEXTABLE
  - select Launch_Site from SPACEXTABLE limit 5
  - select distinct(Landing_Outcome) from SPACEXTABLE
  - select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer == "NASA (CRS)“
  - select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version == "F9 v1.1“
  - select min(Date) from SPACEXTABLE where Landing_Outcome=='Success (ground pad)’
  - select Booster_Version from SPACEXTABLE where (PAYLOAD_MASS__KG_ > 4000) and (PAYLOAD_MASS__KG_ < 6000) and (Mission_Outcome == "Success")
  - select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
- https://github.com/ringo-alin/IBE-DataScience-Assignments/blob/main/Course%2010/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- https://github.com/ringo-alin/IBE-DataScience-Assignments/blob/main/Course%2010/lab_jupyter_launch_site_location.jupyterlite.ipynb

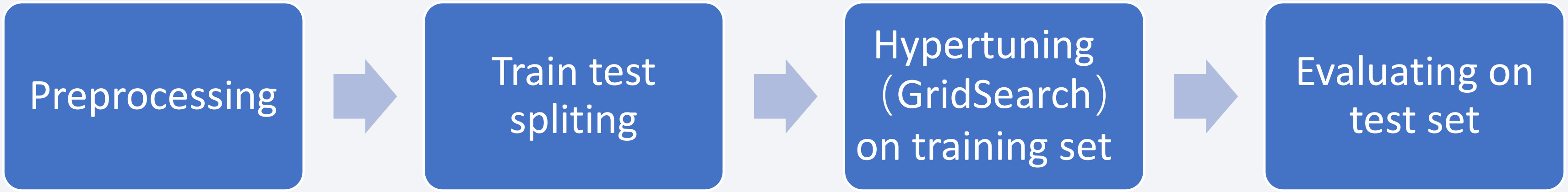
Build a Dashboard with Plotly Dash

- 

- https://github.com/ringo-alin/IBE-DataScience-Assignments/blob/main/Course%2010/spacex_dash_app.py

Predictive Analysis (Classification)

•



- https://github.com/ringo-alin/IBE-DataScience-Assignments/blob/main/Course%2010/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Accuracy on test set

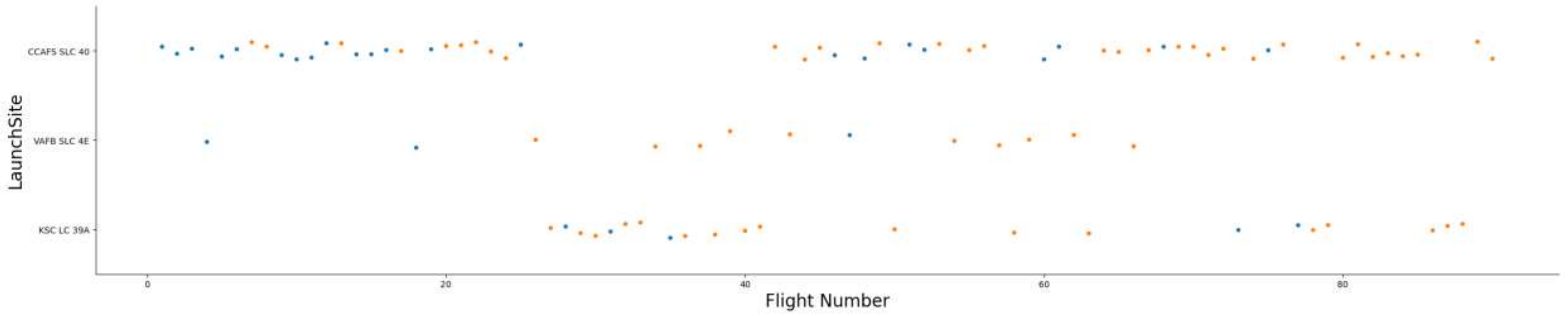
Model	Accuracy
Logistic regression	0.83
SVM	0.83
Decision Tree	0.78
KNN	0.83

Section 2

Insights drawn from EDA

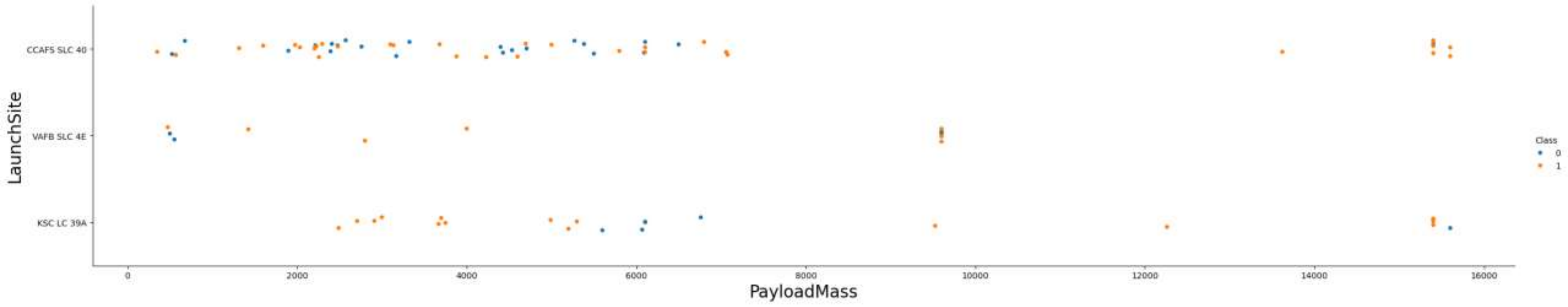
Flight Number vs. Launch Site

- As the flight number increases, the first stage is more likely to land successfully



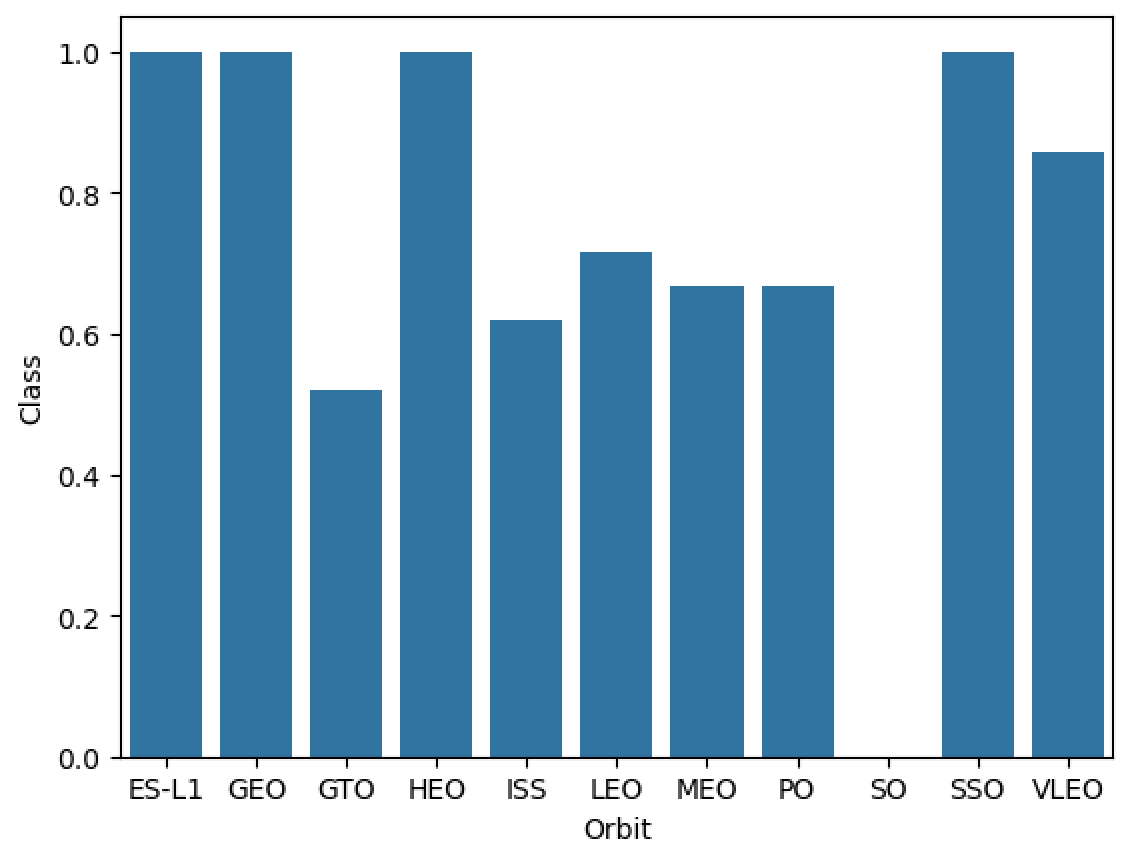
Payload vs. Launch Site

- the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



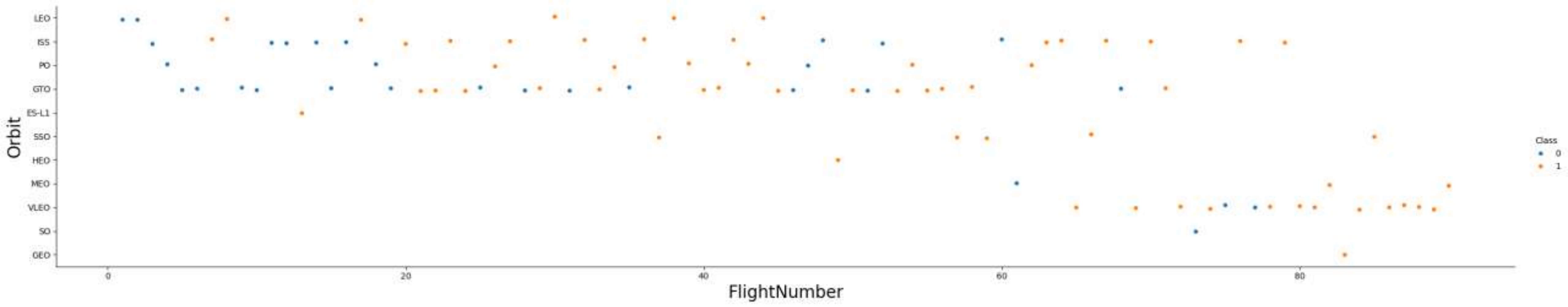
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO have high success rate



Flight Number vs. Orbit Type

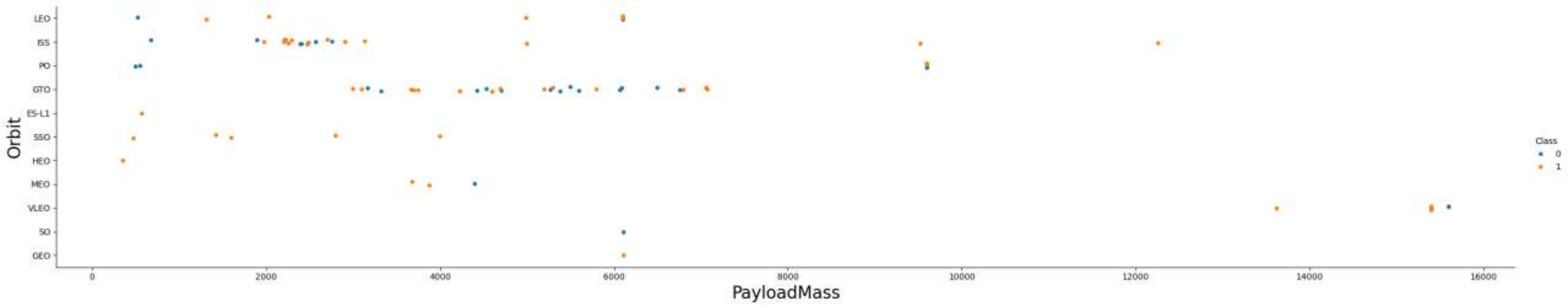
- in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

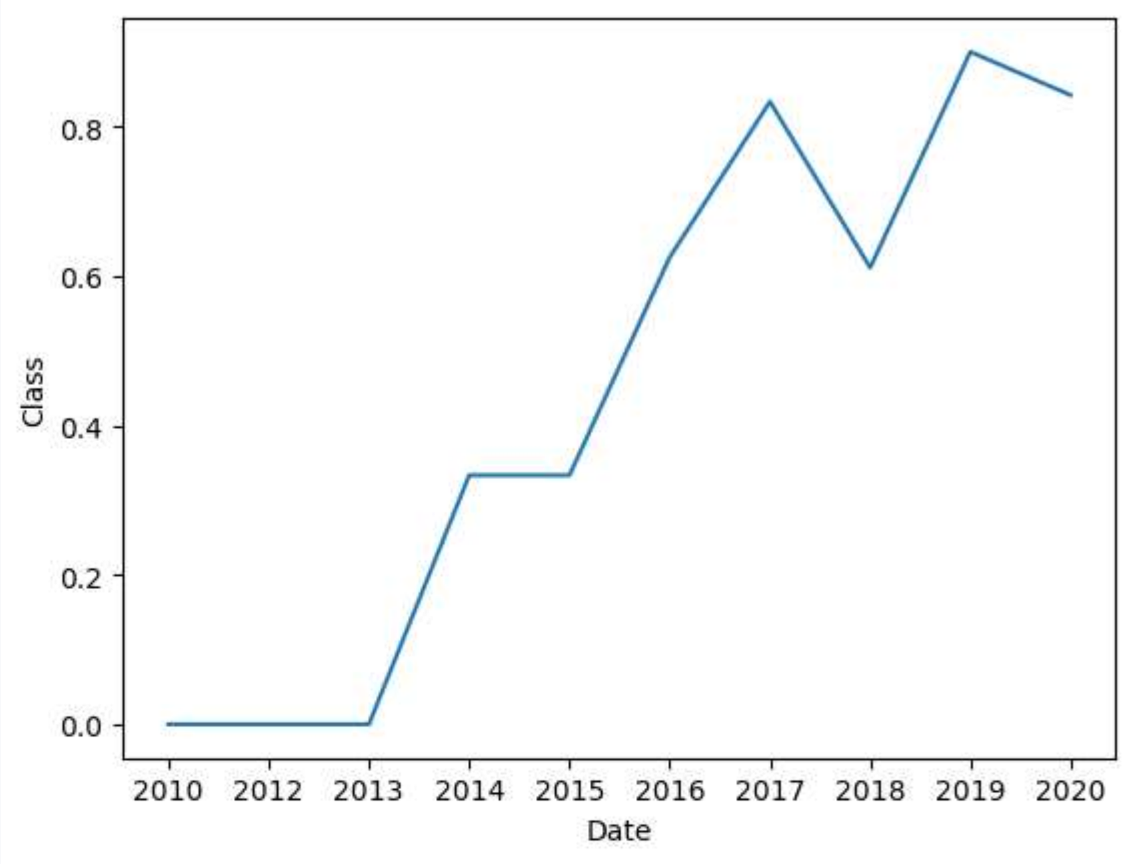
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites

```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%sql select Launch_Site from SPACEXTABLE limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer == "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version == "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome=='Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select Booster_Version from SPACEXTABLE where (PAYLOAD_MASS__KG_ > 4000) and (PAYLOAD_MASS__KG_ < 6000) and (Mission_Outcome == "Success")
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_==(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where (Landing_Outcome =="Failure (drone ship)") and (substr(Date,0,5)=='2015')
```

```
 * sqlite:///my_data1.db
Done.
```

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE where (Date between '2010-06-04' and '2017-03-20') group by Landing_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

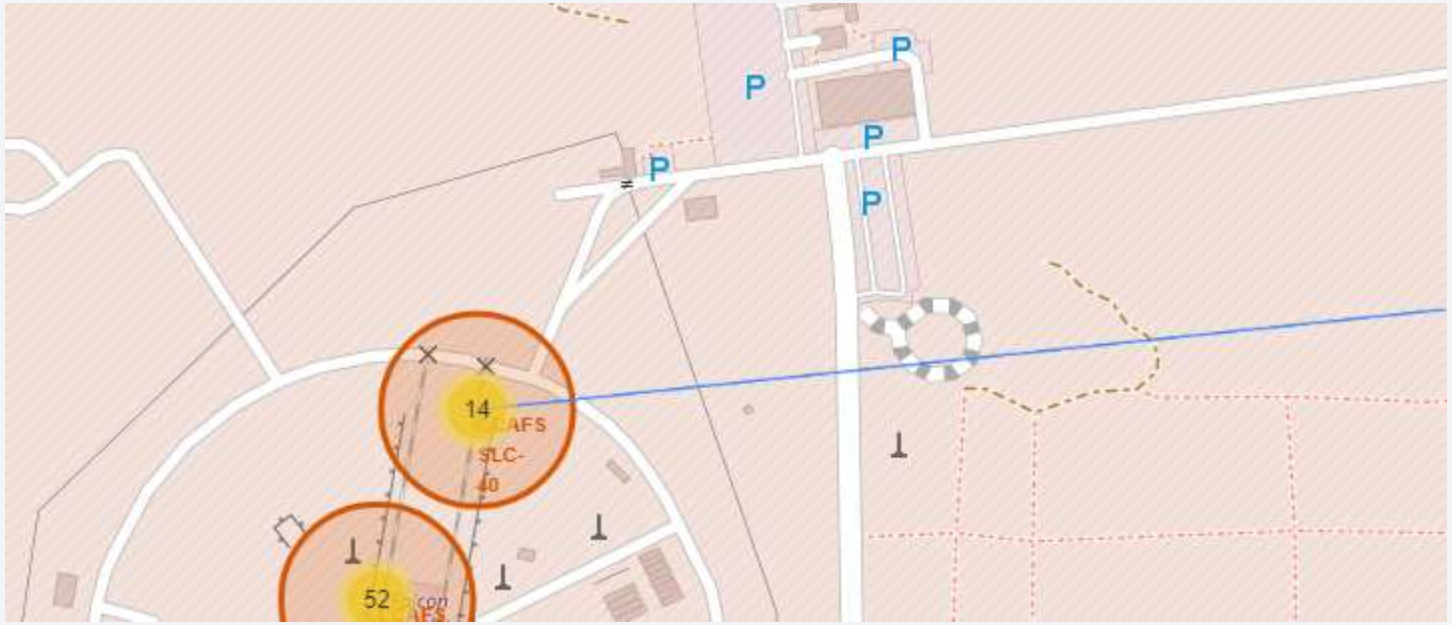
Section 3

Launch Sites Proximities Analysis

Global location view

Zoom in

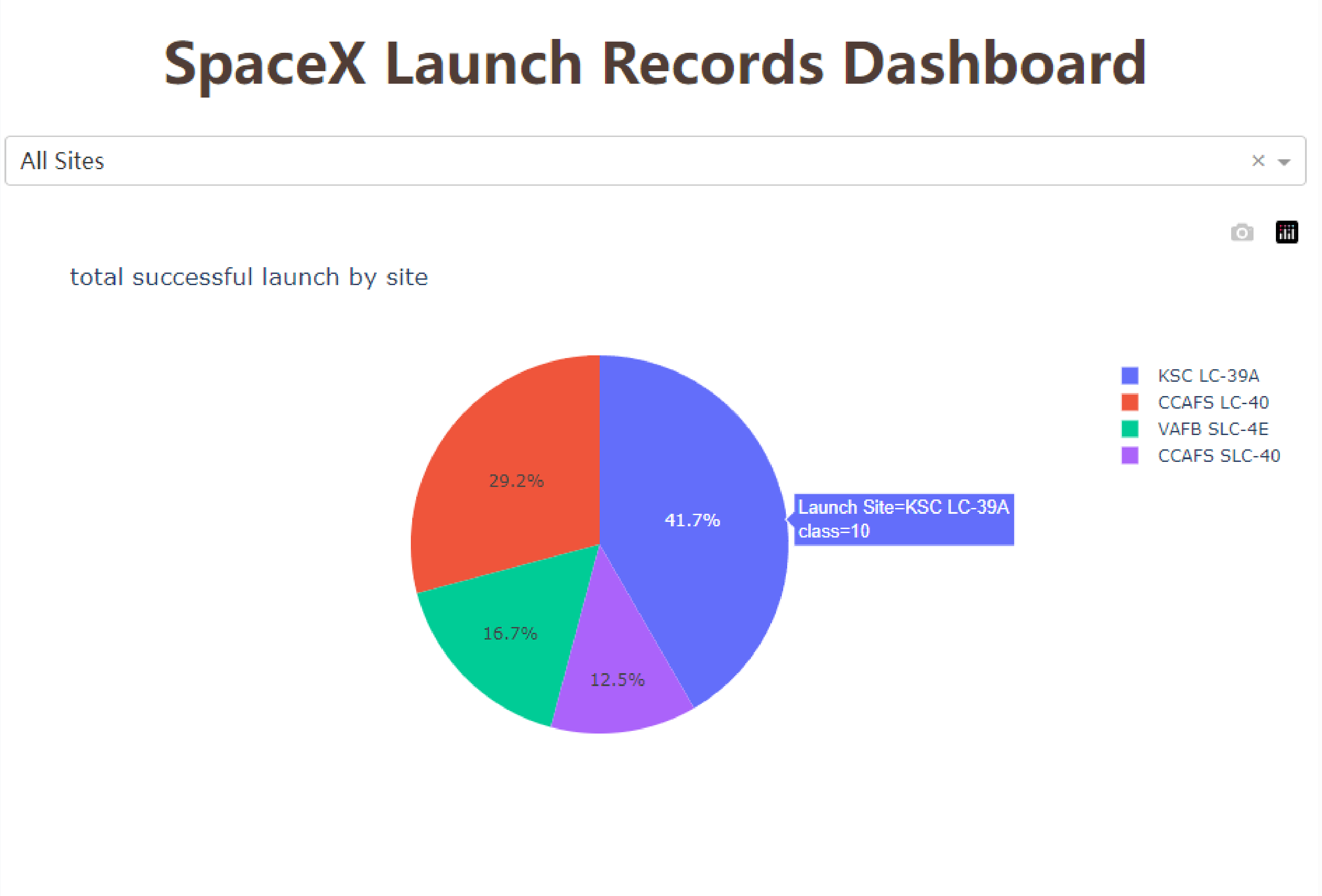
Distance



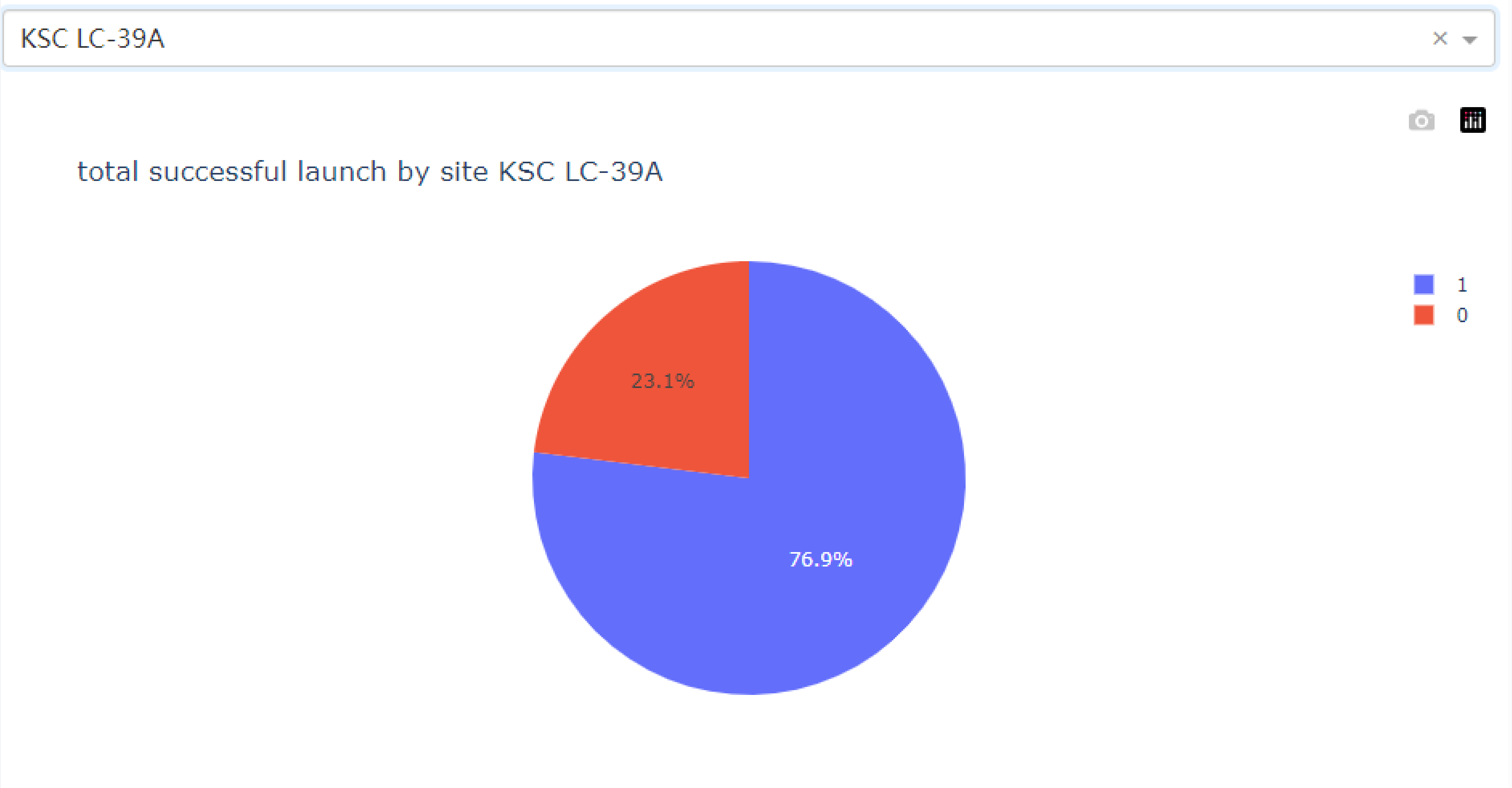
Section 4

Build a Dashboard with Plotly Dash

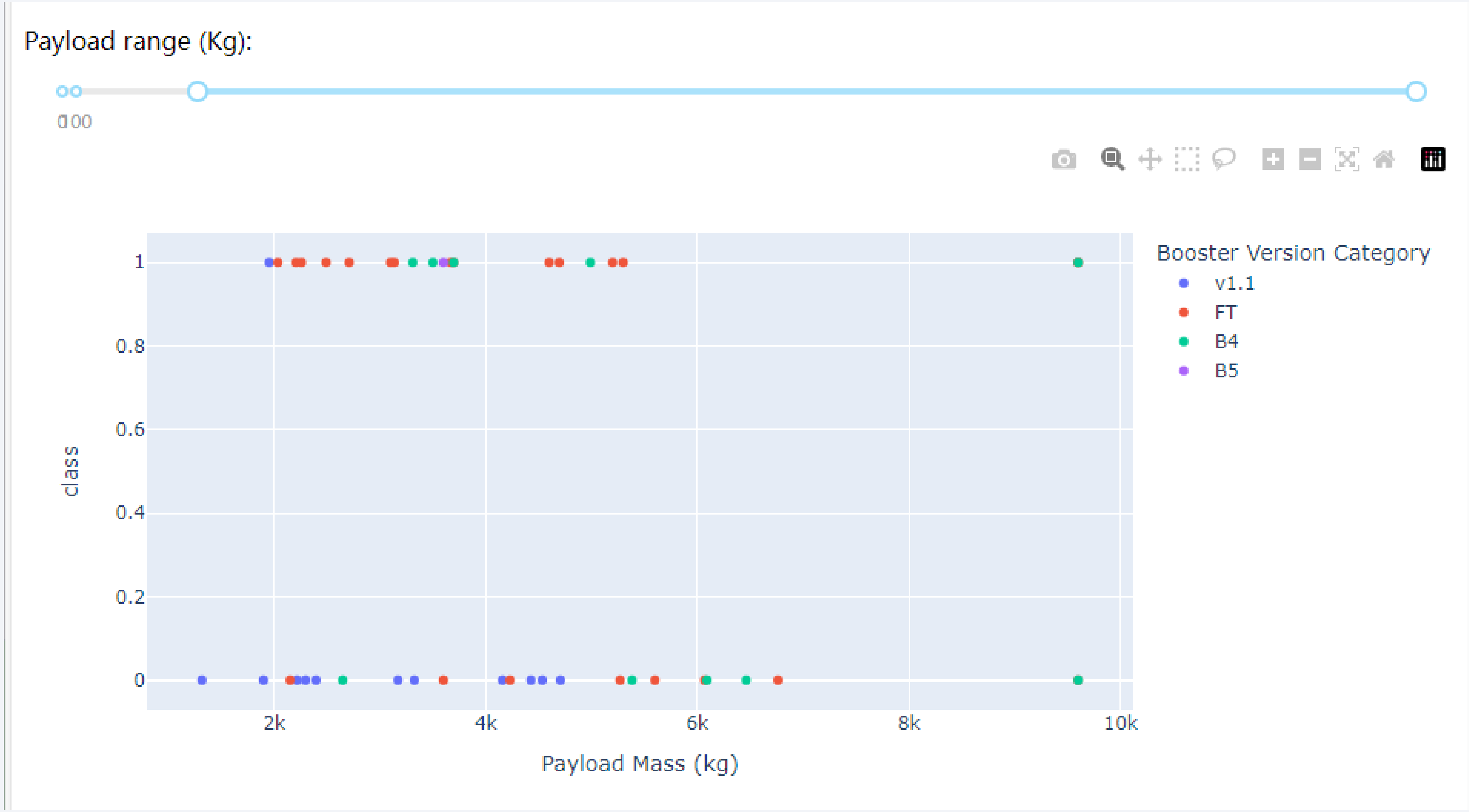
Launch success count for all sites



Success rate by site



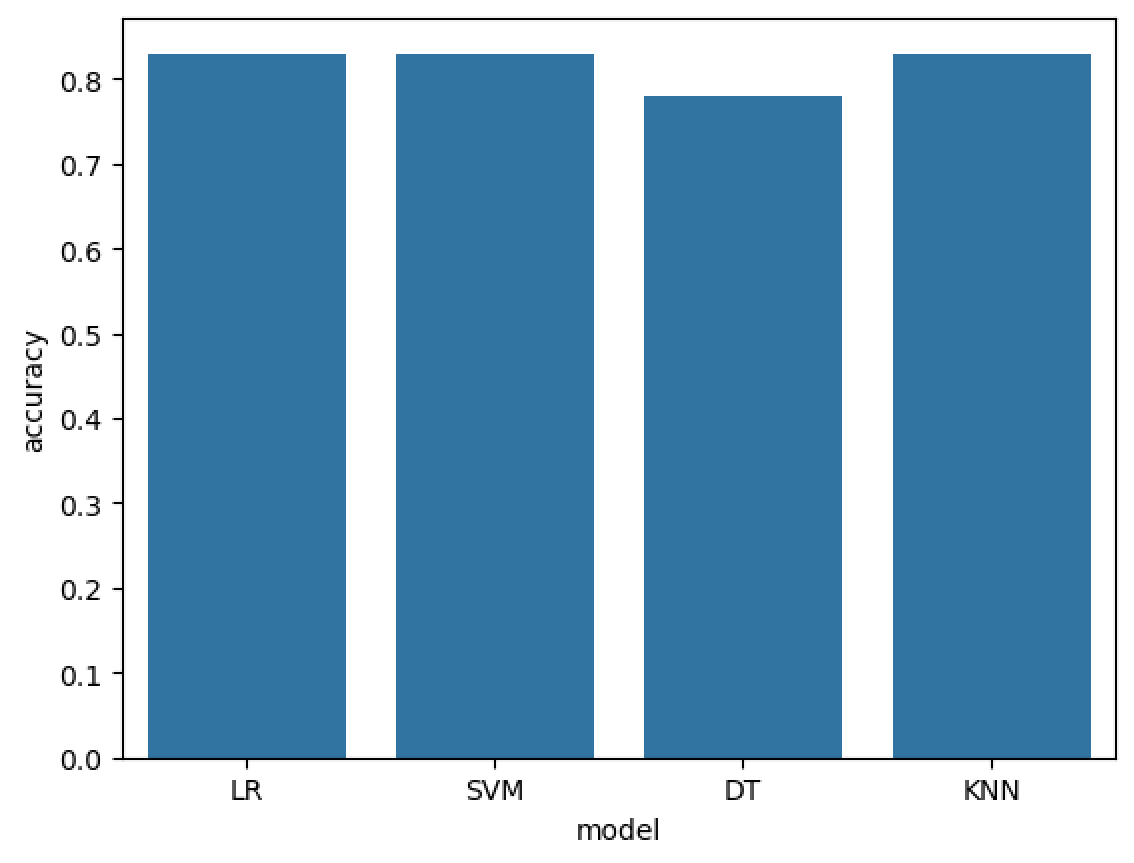
Payload vs launch outcome



Section 5

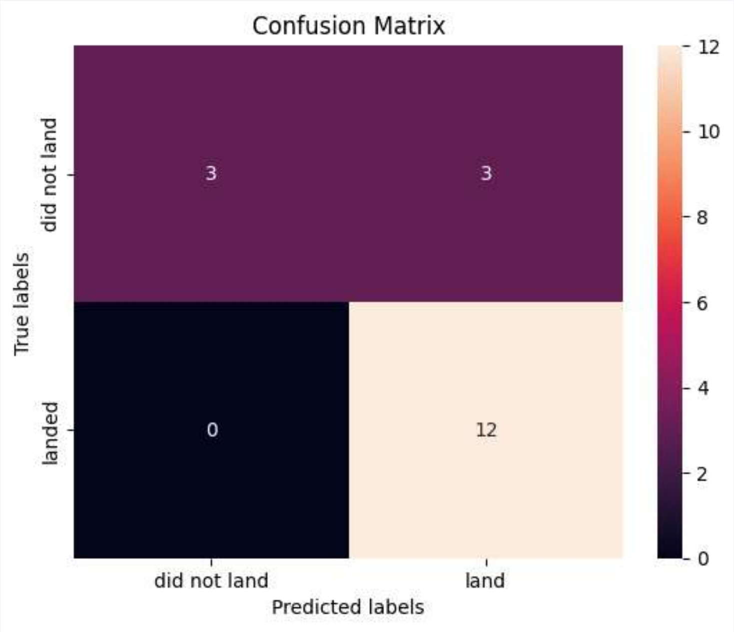
Predictive Analysis (Classification)

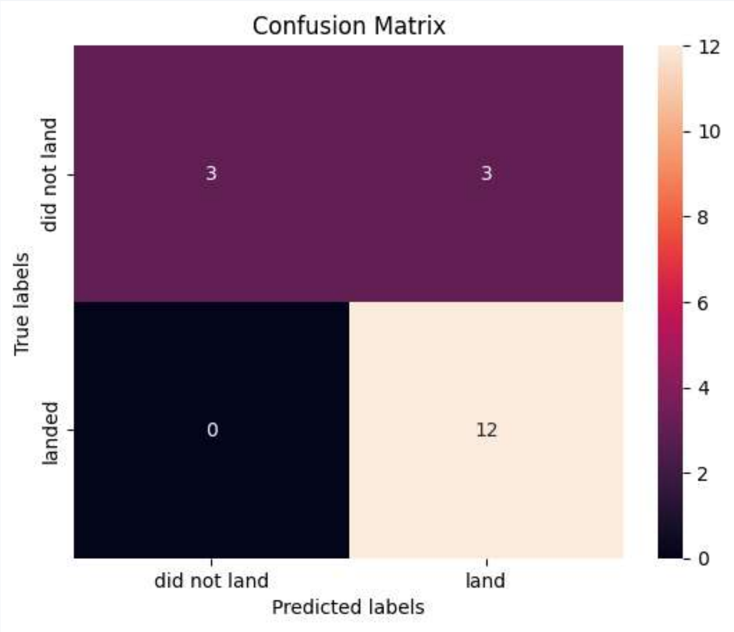
Classification Accuracy

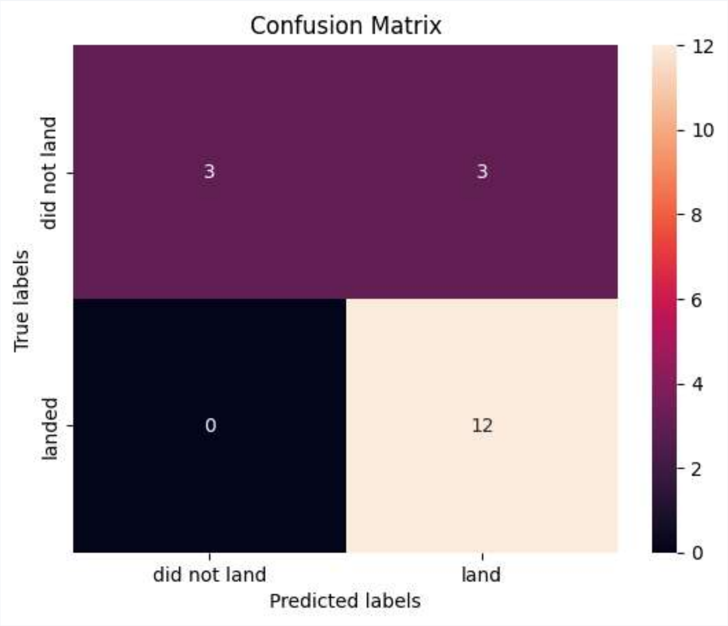


Confusion Matrix

- the confusion matrix of the best performing model







Conclusions

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.

Thank you!