SPACEX FIRST STAGE LANDING ANALYSIS

Misha Freidin

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Summary of methodologies

- Data Collection with Beautiful Soup and SpaceX API
- Data Wrangling with Pandas
- EDA using SQL, visualisation with Dash, Seaborn and Matplotlib

INTRODUCTION

- The problem is to predict whether the first stage of a Falcon 9 launch will be reused. This determines the ultimate cost of a launch
- We use the dataset to determine which factors affect whether stage 1 was successfully recovered

METHODOLOGY

- Data Collection: REST API and Web-scraping Wikipedia with BeautifulSoup
- Data Wrangling
- Exploratory Data Analysis (EDA) - Visualisation
- Exploratory Data Analysis (EDA) - Databases
- Visual Analysis using Folium
- Visual Analysis using Dash by Plotly
- Machine Learning Classification Modelling using Scikit-Learn

DATA COLLECTION: REST API AND WEB-SCRAPING WIKIPEDIA WITH BEAUTIFULSOUP

- https://github.com/mishadcf/Data-Science-Course/blob/main/jupyter-labs-spacex-data-collection-api-4.ipynb: Complete code available here
- Overview:
- Call the SpaceX API using requests
- Create and fill a Pandas dataframe using the API responses, normalizing JSON file
- Drop the records pertaining to Falcon 1 launches
- Update the dataframe with relevant columns
- Find missing values and replace with the mean of the attribute (s)

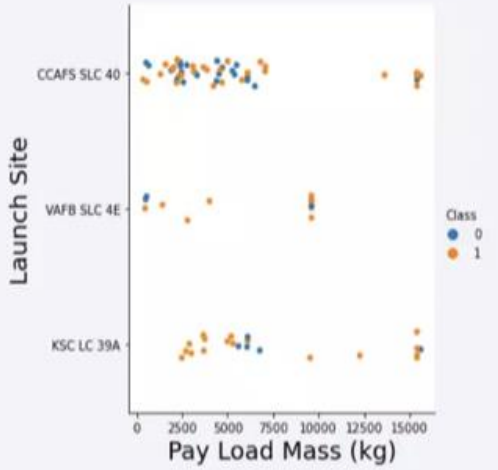
DATA WRANGLING

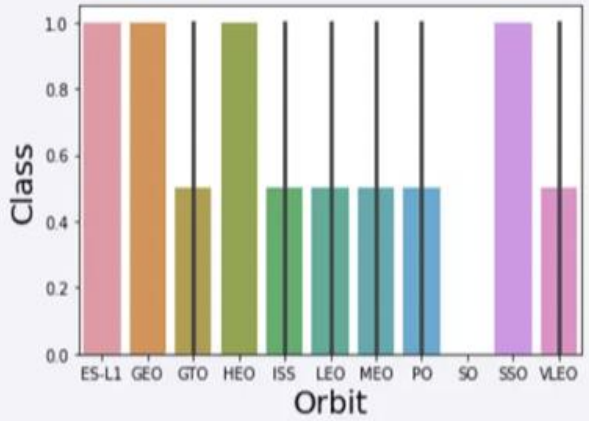
- https://github.com/mishadcf/Data-Science-Course/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb
- Overview:
- As seen in the github repo:
- We analysed the data to determine which training labels to use in our machine learning models
- We created a landing_outcome column to assign a numerical value to the success of a launch

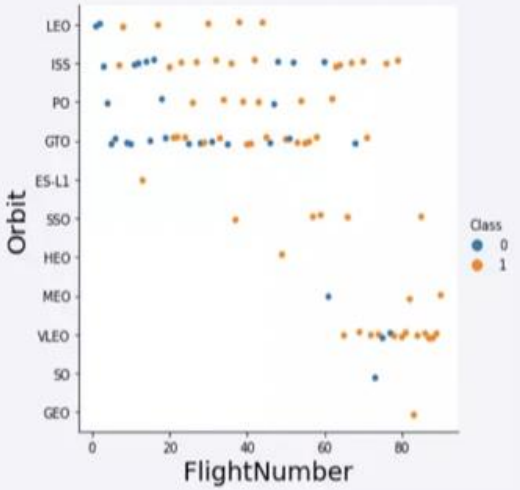
EXPLORATORY DATA ANALYSIS (EDA) - VISUALISATION

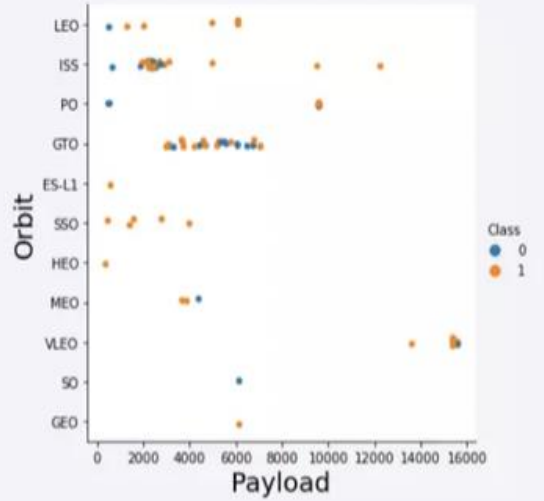
- https://github.com/mishadcf/Data-Science-Course/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite-2.ipynb
- Using Seaborn, Matplotlib; we explored the relationships between FlightNumber and Payload variables and the target variable: launch_outcome

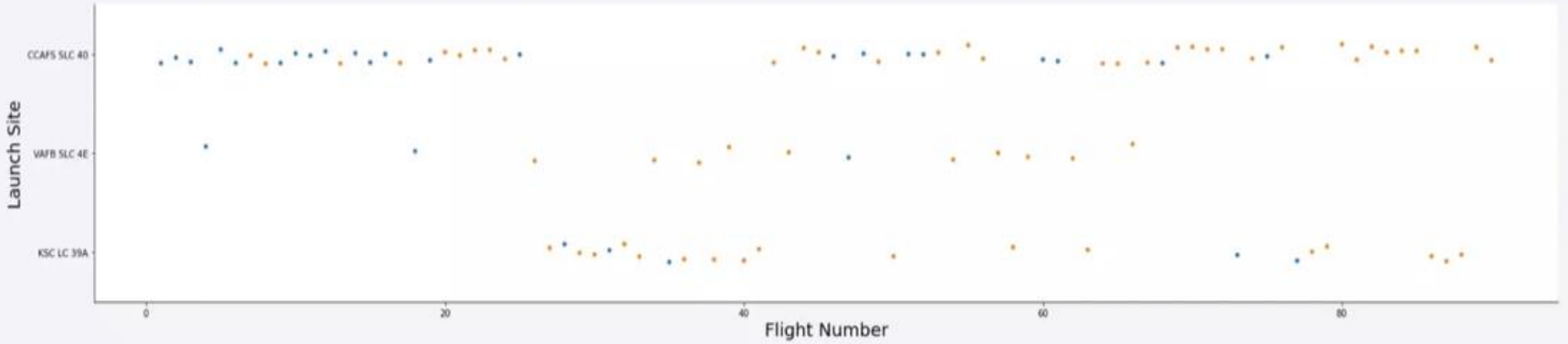
(EDA) - VISUALISATION GRAPHICS

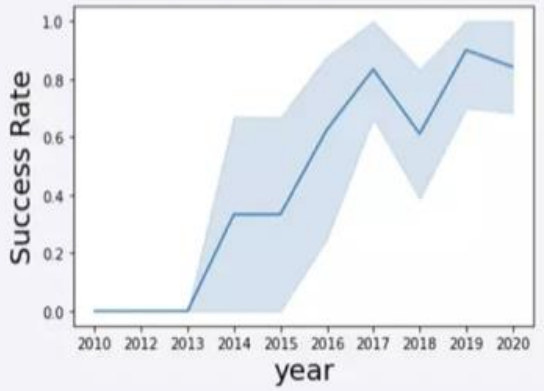










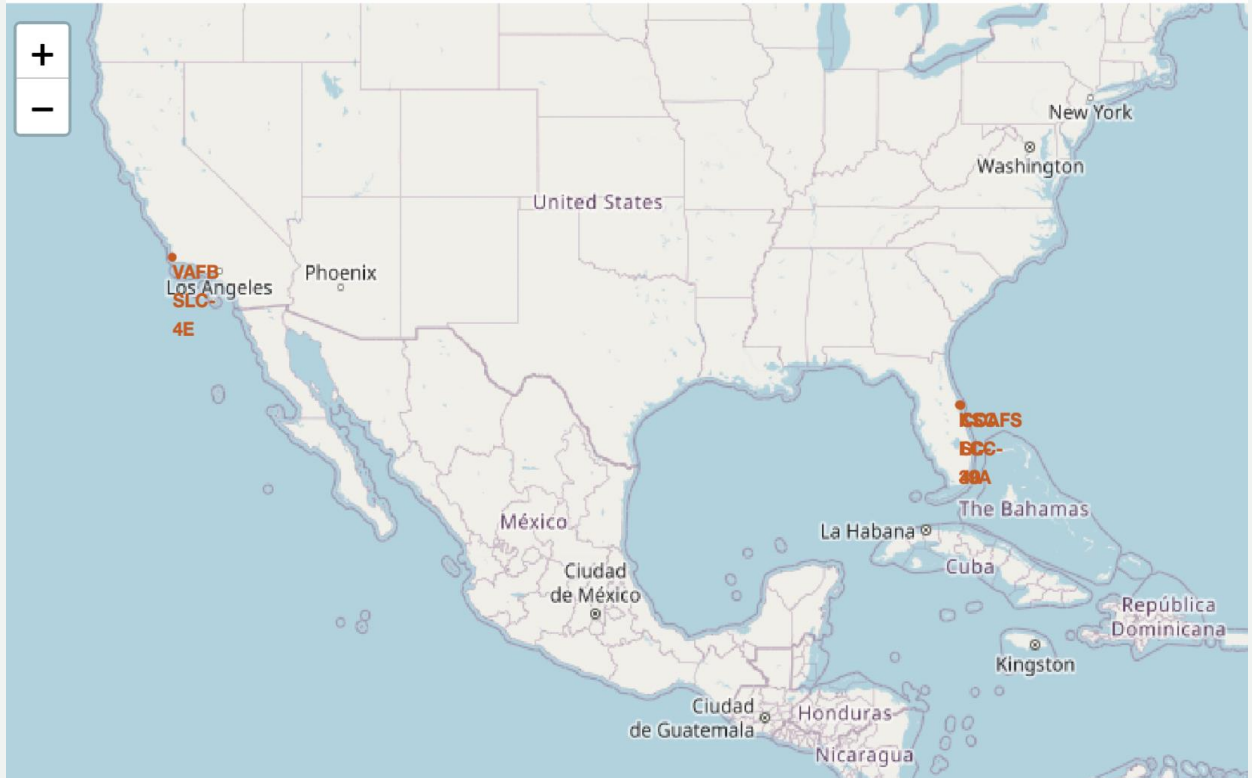


EXPLORATORY DATA ANALYSIS (EDA) - DATABASES

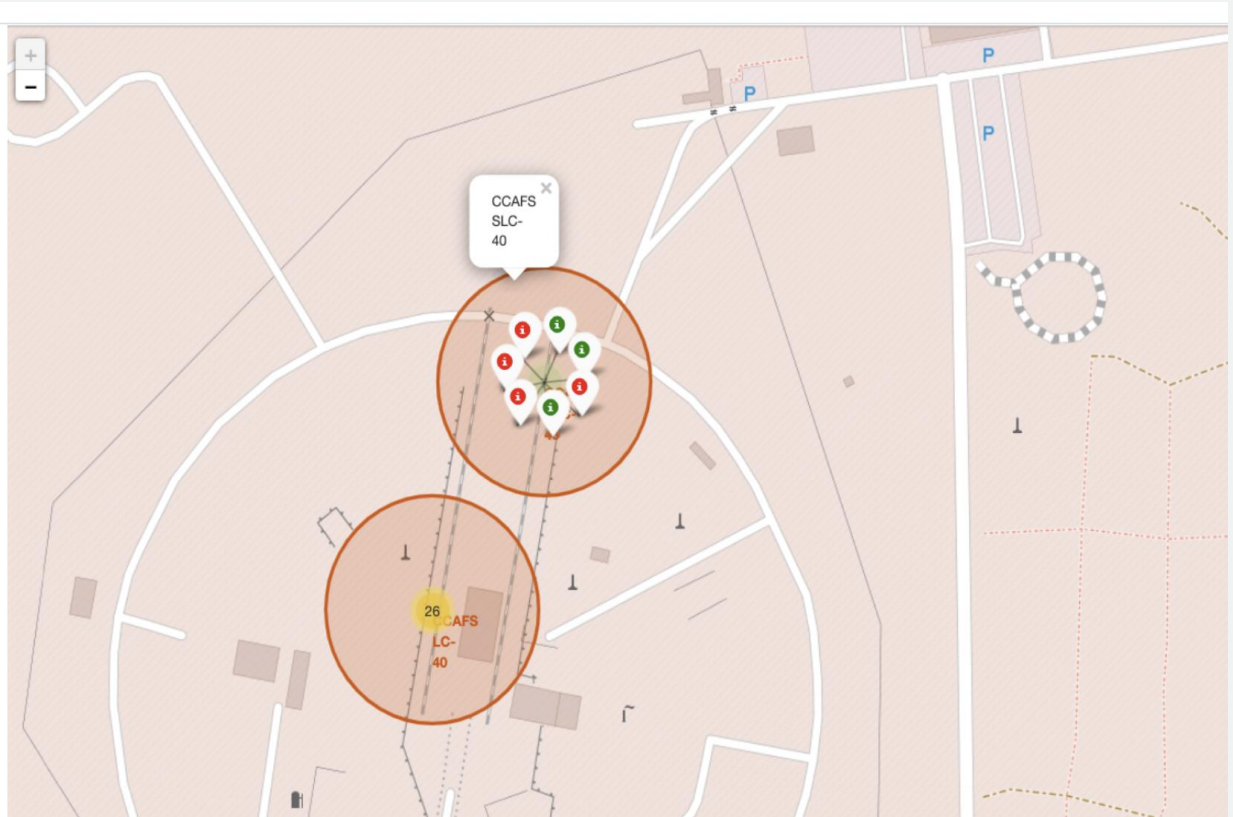
- Using SQL (as documented here:),
- https://github.com/mishadcf/Data-Science-Course/blob/main/jupyter-labs-eda-sql-coursera_sqllite-2.ipynb
- We answered among the following:
  - What are the unique launch site names?
  - What was the total payload mass carried by boosters launched by NASA(CRS)
  - The first date where a ground pad landing happened succesfully
  - The total failed and succesful mission outcomes
  - The booster versions carrying the maximum payloads
  - The months and records of failed drone ship landings
- The count of landing outcomes grouped by outcome type between 04-06-2010 and 20-03-2017

VISUAL ANALYSIS USING FOLIUM

Creating a folium map object and adding markers in for each launch site

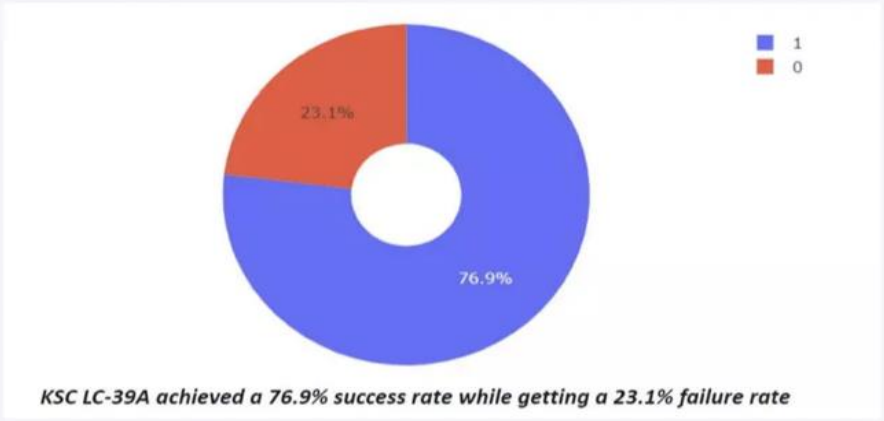


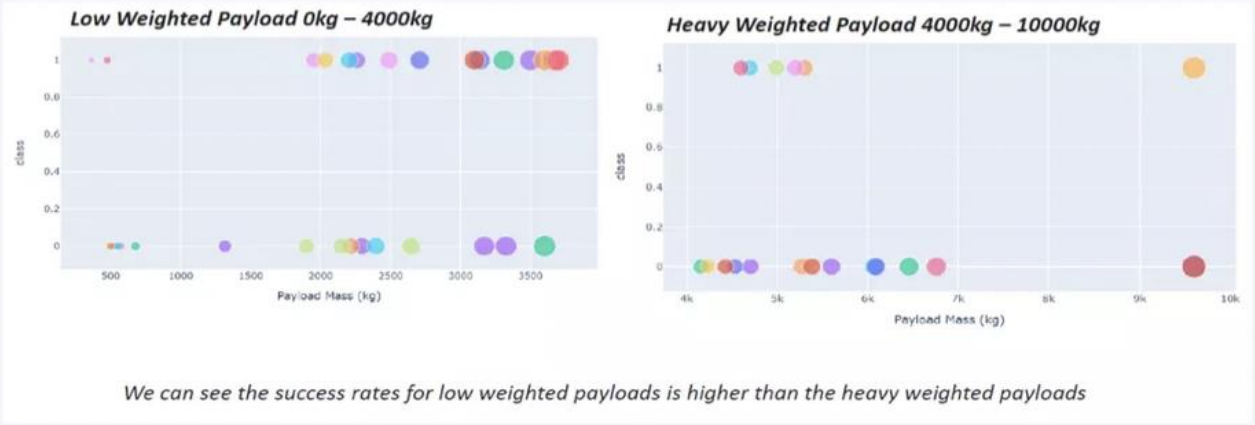
Using a for loop, we represent launch outcomes as different color clusters



VISUAL ANALYSIS USING DASH BY PLOTLY

Below are some of the most interactive graphs we generated with Dash



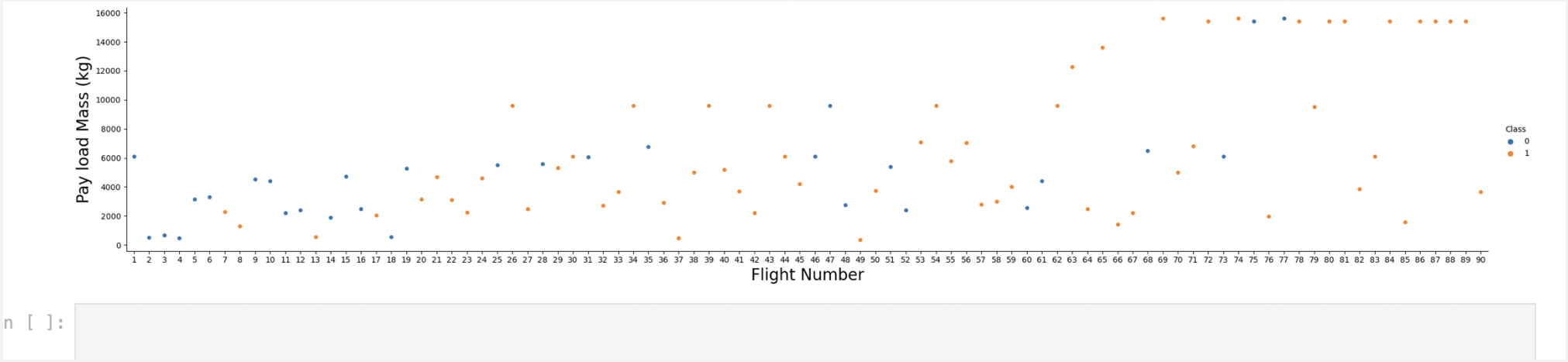


ML CLASSIFICATION MODELLING USING SCIKIT-LEARN

RESULTS

- EDA Pandas/ Matplotlib Code: https://github.com/mishadcf/Data-Science-Course/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite-3.ipynb
- SQL EDA Code: https://github.com/mishadcf/Data-Science-Course/blob/main/jupyter-labs-eda-sql-coursera_sqllite-2.ipynb

EDA PANDAS/MATPLOTLIB RESULTS

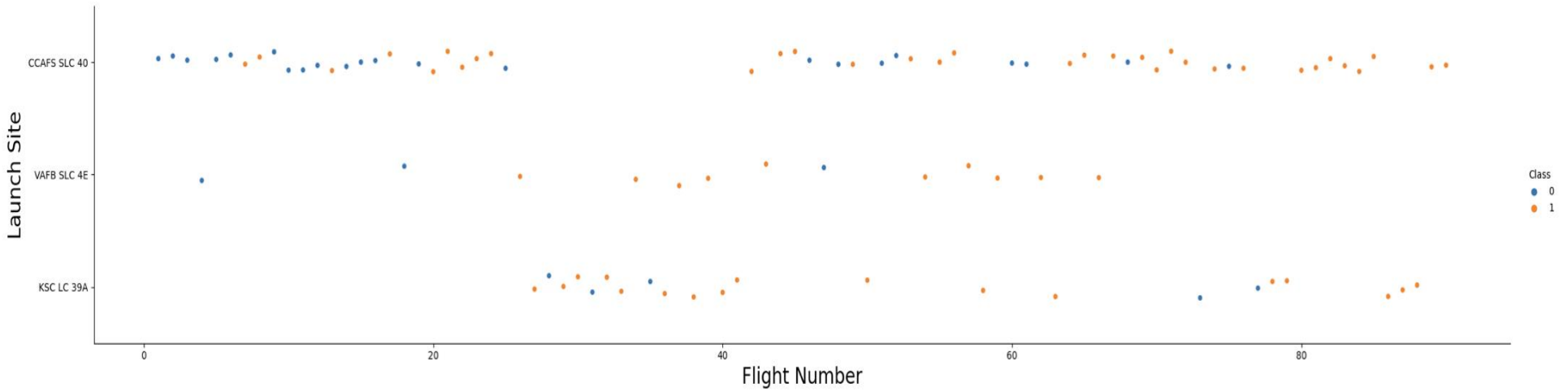


n [ ]:

- From our plot above:

- Payload mass average is increasing over flight number
- The variability of payload mass is increasing over time (likely due to the increased choice of payload, coming from technological advance. Clearly, if you could send a lower payload before, with new tech, you still have the former option. Now you have the choice of each.
- The Class denotes whether the launches were able to land the first stage (our target variable for the prediction)
- With a greater proportion of orange dots to blue, with flight number increasing- we can see the success rate of stage 1 landings are improving

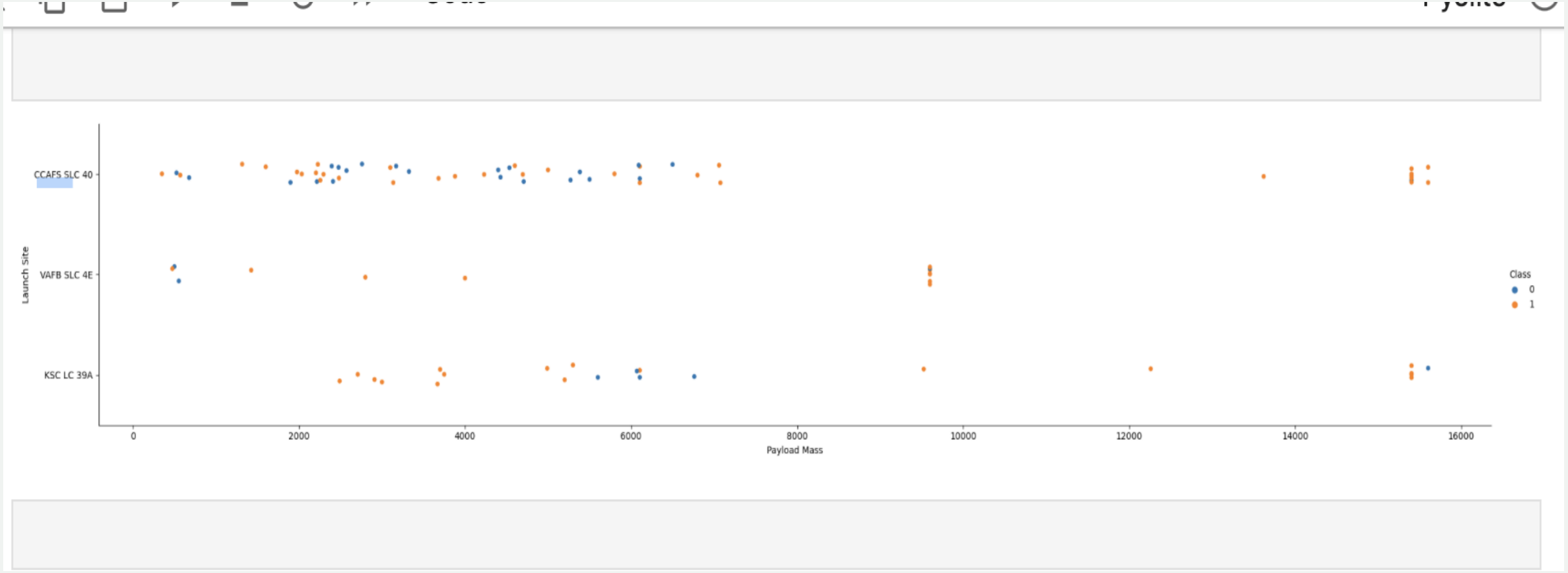
EDA PANDAS/MATPLOTLIB RESULTS 2



Insights

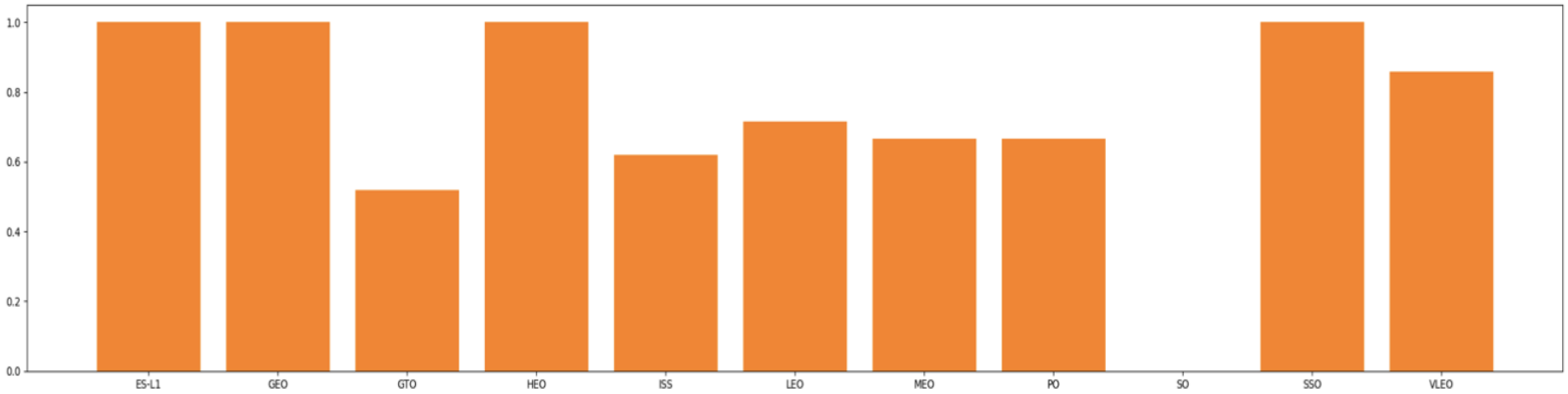
- CCAFS SLC-40 has the most launches
- There aren't flights for VAFB SLC-4E after approx flight number 40

EDA PANDAS/MATPLOTLIB RESULTS



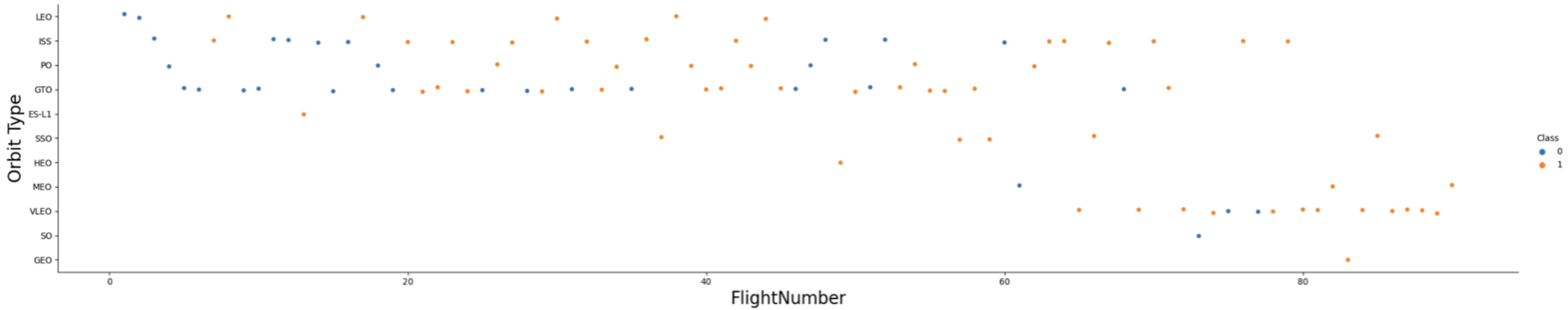
- https://github.com/mishadcf/Data-Science-Course/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite-3.ipynb

EDA PANDAS/MATPLOTLIB RESULTS

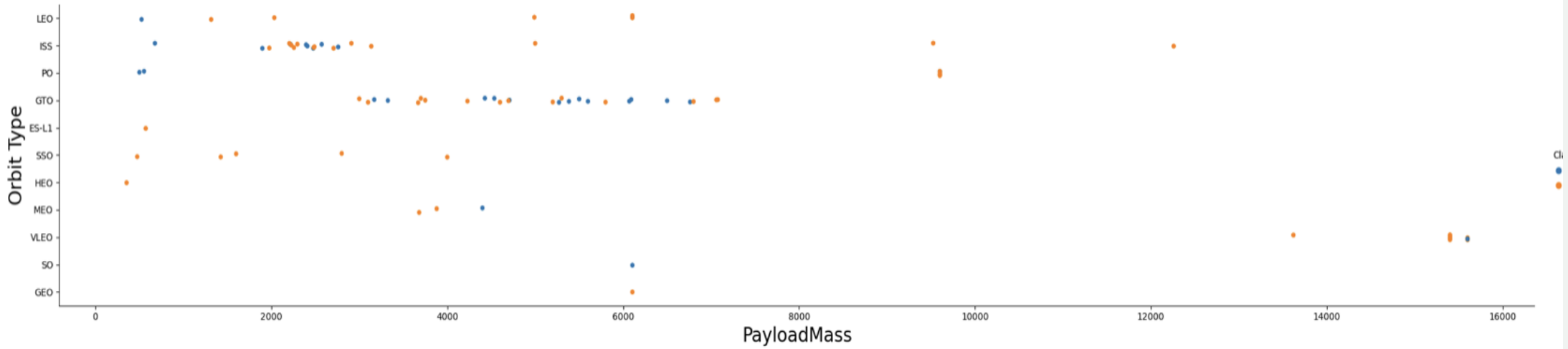


ES-L1, GEO, HEO, SSO have the best success rates

FLIGHT ORBIT TREND



```
plt.show()
```


LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
Orbit Type
0
2000
4000
6000
8000
10000
12000
14000
16000
PayloadMass

FEATURE SELECTION

- We changed the categorical features into dummy variables
- Used one-hot-encoding to prepare the features for the machine learning models

EDA SQL RESULTS

- After querying the data, the most important insights are:
- There are 4 distinct launch sites: CCAFS-LC 40, CCAFS-SLC 40, KSC LC-39A, VAFB SLC-4E
- NASA CRS launched a total of 46 tonnes of payload mass in our dataset
- The first successful ground pad landing was recorded on 01-05-2012
- https://github.com/mishadcf/Data-Science-Course/blob/main/jupyter-labs-eda-sql-coursera_sqllite-2.ipynb (Here you can see all of the queries)

EDA SQL RESULTS

t[47]:

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

The distinct launch sites of the Falcon 9

EDA SQL RESULTS

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

t[42]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

n [ ]:

The first five records of launches from CCAFS sites

EDA SQL RESULTS

```
In [30]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[30]:
```

sum(PAYLOAD_MASS__KG_)
45596

The total mass (KG) of payload sent by NASA (CRS) by Falcon 9 over all launches

```
In [ ]:
```

```
In [35]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
Out[35]:
```

avg(PAYLOAD_MASS__KG_)
2928.4

Average Payload Mass (KG) sent from rockets with booster version f9 v.1.1

EDA SQL RESULTS

```
%sql SELECT min(date) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)'
```

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

min(date)
01-05-2017

An important date for SpaceX:
on 01-05-2017, the first Falcon 9 Ground Pad Landing was succesful

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

names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA SQL RESULTS

Done.

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

total number of successful and failed mission outcomes

[64]:

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

Booster Versions that carried the maximum payloads

EDA SQL RESULTS

Done.

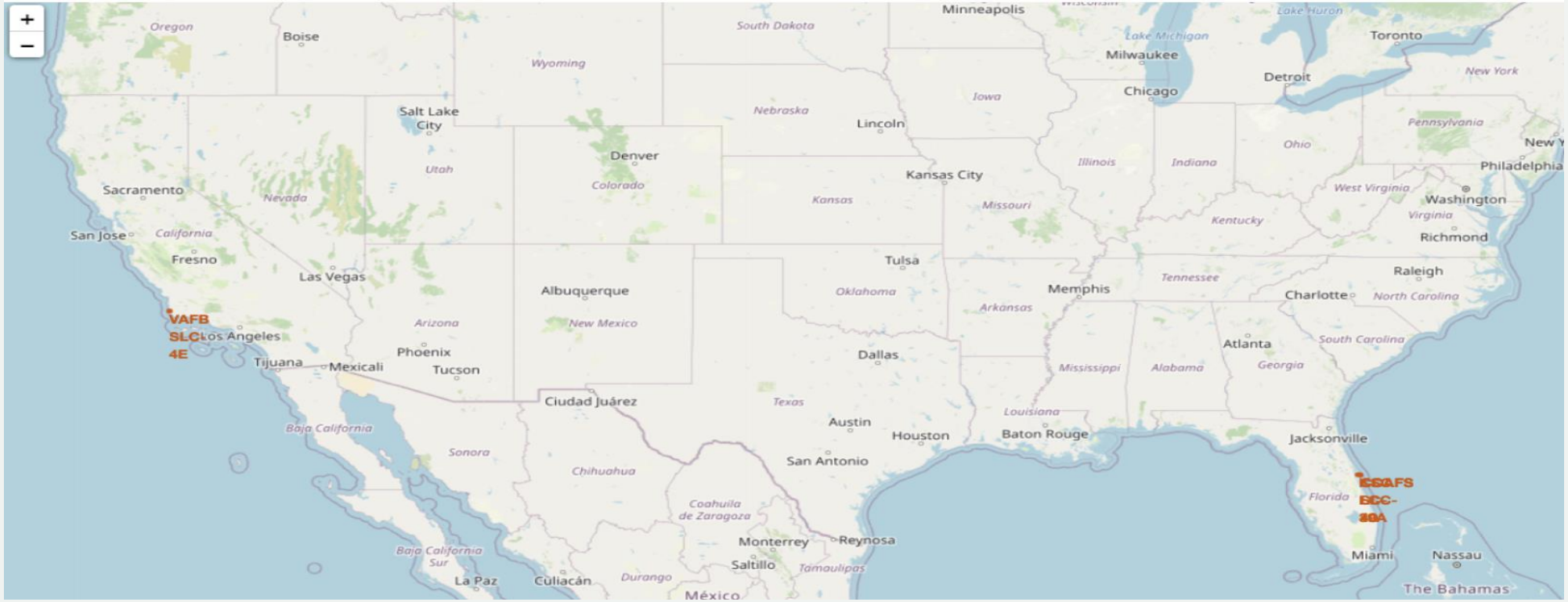
month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Details of failed drone ship landings in 2015

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

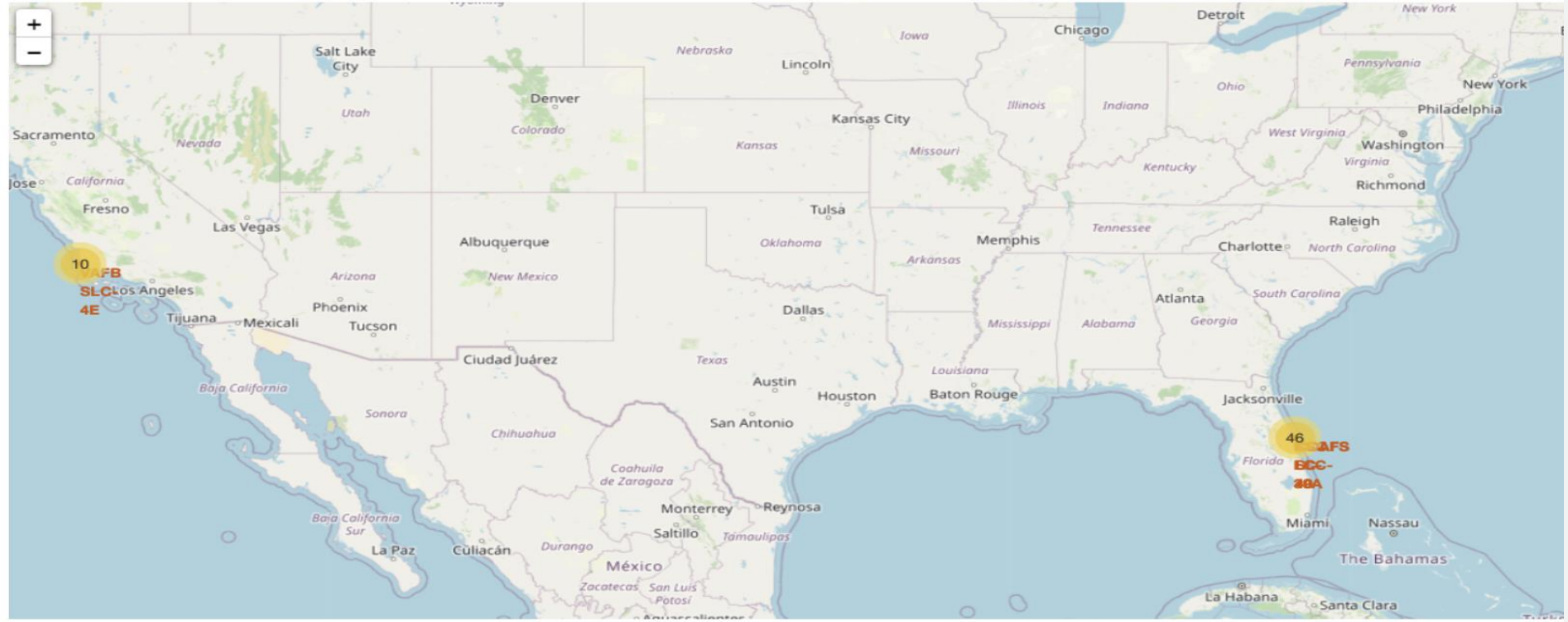
Landing Outcomes count

RESULTS OF USING FOLIUM



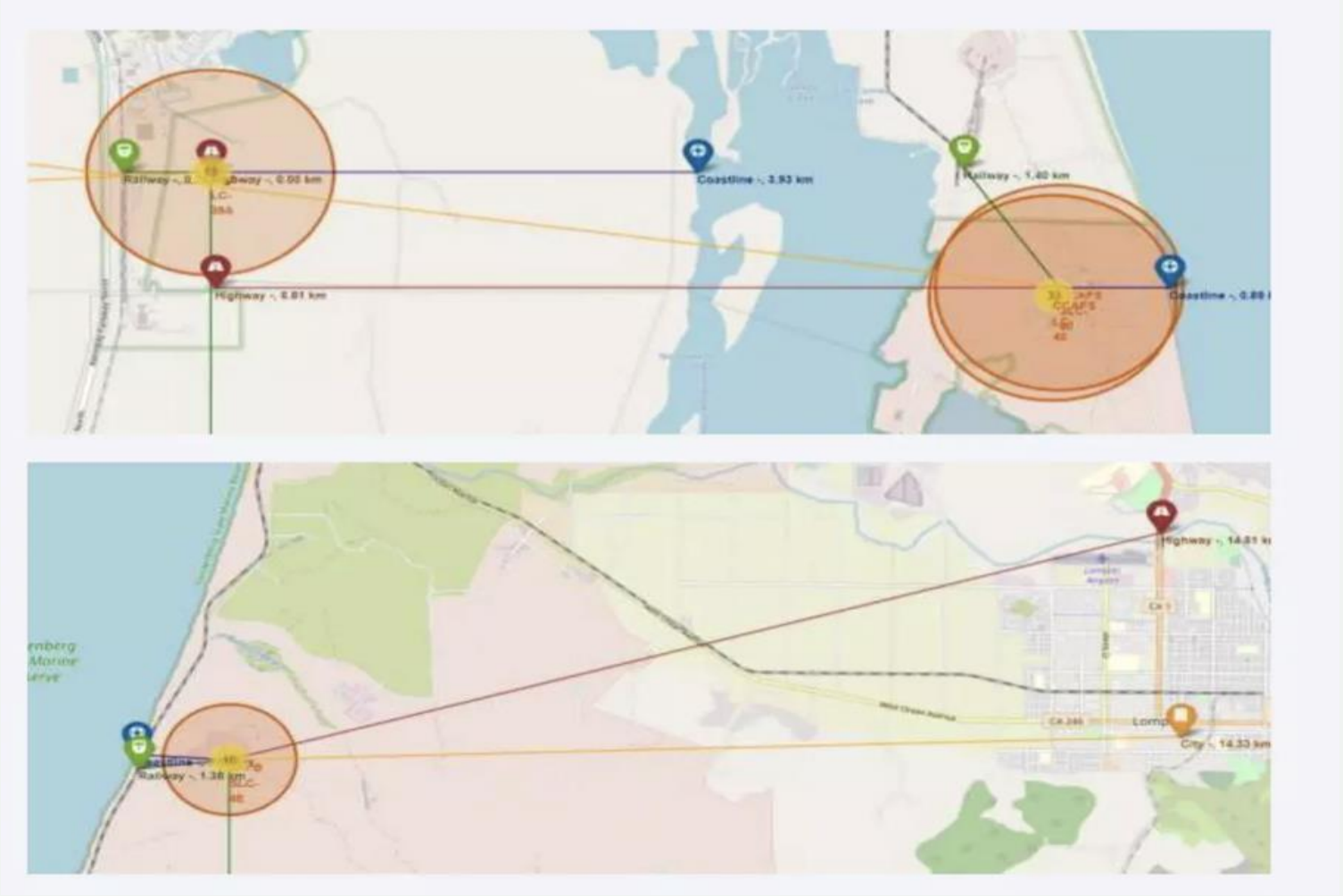
The different launch sites of the Falcon 9, all coastal, concentrating on 2 regions

RESULTS OF USING FOLIUM



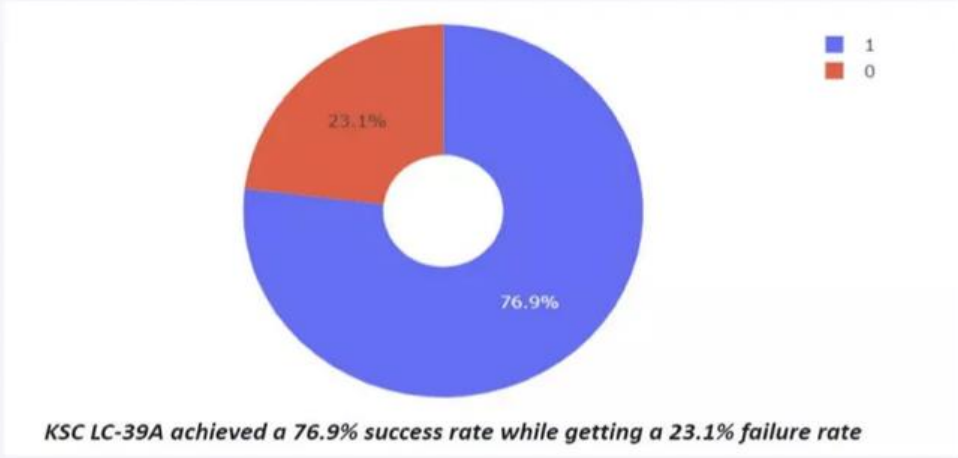
Number of launches plotted

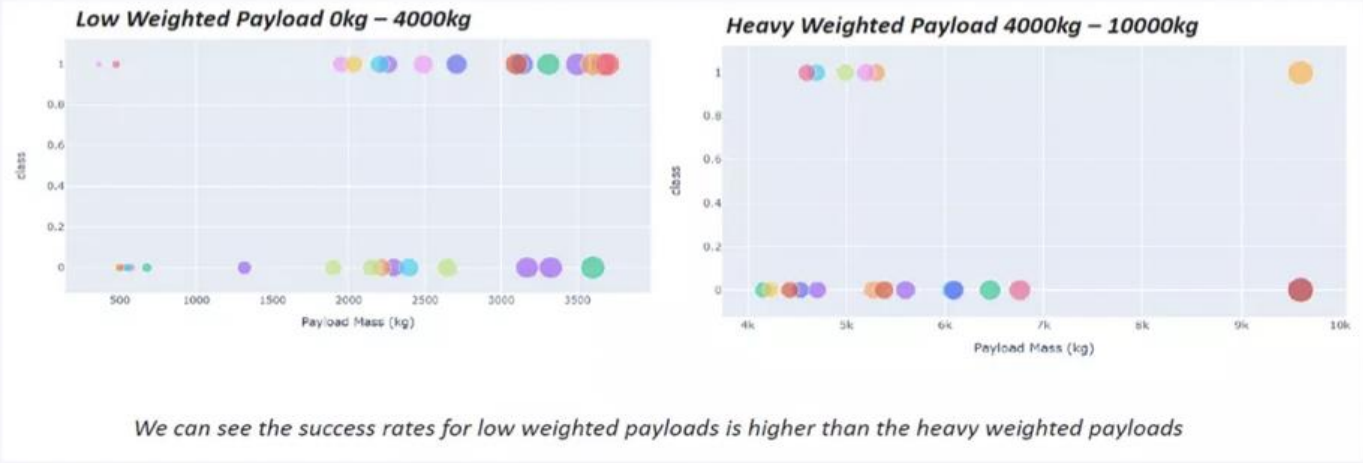
RESULTS OF USING FOLIUM



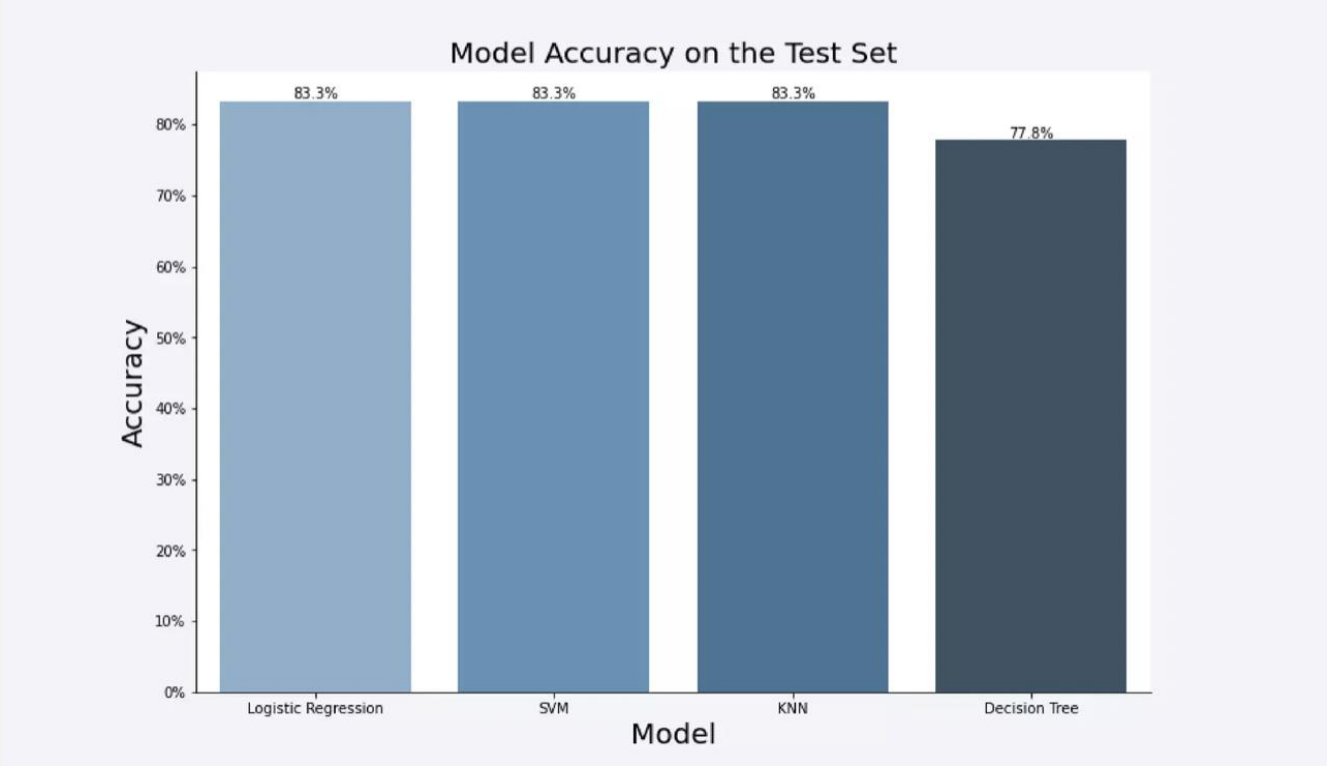
Distance between each site and the nearest railway, coastline and highway

DASH RESULTS

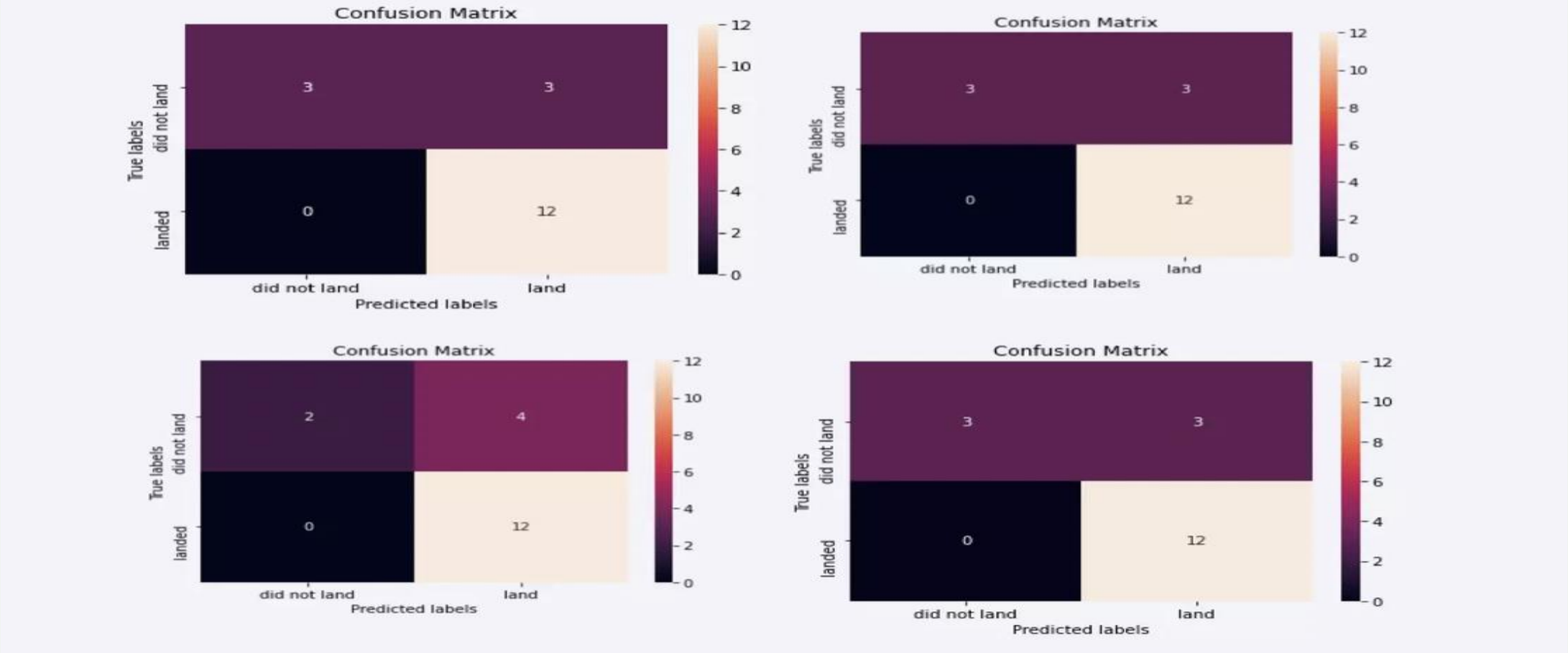




ML CLASSIFICATION RESULTS BY MODEL TYPE



ML CLASSIFICATION RESULTS CONFUSION MATRICES



MODEL VALIDATION

- Our analysis suggests Logistic Regression, K Nearest Neighbours and Support Vector Machines are the most accurate, from our test set
- More data may be needed to accurately predict the landing outcomes, we only have 90 data points
- Logistic Regression, KNN, SVM have 83.3% accuracy on our test set, whereas Decision trees give a 77.8% accuracy

CONCLUSION

- VAFB SLC-4E has the highest success rate of Falcon 9 launch sites
- The success rate of launches is increasing with time
- The orbits with the greatest success rates are ES-L1, GEO, HEO, SSO
- For machine learning classification models Logistic Regression, K Nearest Neighbours and Support Vector Machines are the most accurate according to our validation procedure
- More data is required to be able to accurately predict Falcon 9 outcomes