Data Science Capstone Project

Niranjan N

23 August 2021

Outline

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

Executive Summary

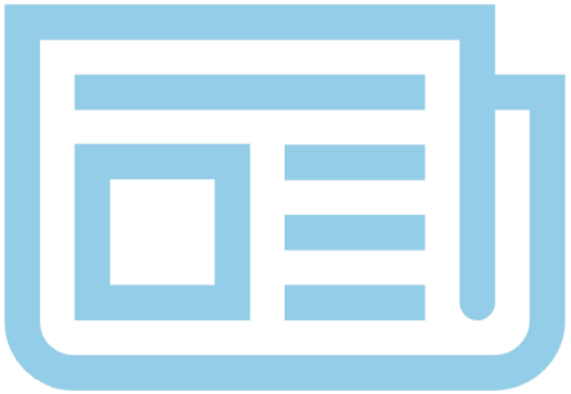
- Summary of methodologies
  - Many methodologies were used to predict whether a SpaceX rocket would land successfully, including data collection, wrangling and visualization, logarithmic regression and machine learning
- Summary of all results
  - The best performing machine learning algorithm is the decision tree classifier and the rocket most likely to land are those with lower-weighted payloads in the orbits of GEO, HEO, SSO and ES-L1

Introduction

- Project background and context
  - SpaceX is an innovative company seeking to democratize space travel through competitive pricing and the reuse of certain launch materials. This project aims to help predict which types of rockets will land successfully. Since SpaceX spends $62 million per launch (compared to $162 million from competitors), accurate predictions will help further drive costs down, eventually making space travel more accessible to more people and companies

- Problems you want to find answers
  - The problems the research set out to answer were:
    - What factors have the most weight in determining whether a rocket will land successfully or not
    - Which machine learning algorithms perform the best in helping us make those predictions

Methodology



- Data collection methodology:
  - Data was collected via the SpaceX Rest API and also via web scraping from wikipedia
- Perform data wrangling
  - Data was processed using One Hot Coding fields for machine learning. Data was also standardized, Means were added to rows without values and Irrelevant data columns were removed from the data set
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - These included evaluating model type and performance

Methodology

Data collection

- Describe how data sets were collected.
  - Data sets were collected by using the GET command to collect and parse json files from the SpaceX API. Then a dataset was created combining data from different tables

- You need to present your data collection process use key phrases and flowcharts

Data collection – SpaceX API

[DataScienceCapstone/Data Collection API Notebook(1).ipynb at main · Nirin13/DataScienceCapstone (github.com)]

```
In [16]: spacex_url="https://api.spacexdata.com/v4/launches/past"

In [17]: response = requests.get(spacex_url)
```

GET request from SpaceX API

```
In [21]: from pandas.io.json import json_normalize
         A = response.json()
         data = json_normalize(A)
```

Normalize data

```
In [30]: launch_dict = {'FlightNumber': list(data['flight_number']),
         'Date': list(data['date']),
         'BoosterVersion':BoosterVersion,
         'PayloadMass':PayloadMass,
         'Orbit':Orbit,
         'LaunchSite':LaunchSite,
         'Outcome':Outcome,
         'Flights':Flights,
         'GridFins':GridFins,
         'Reused':Reused,
         'Legs':Legs,
         'LandingPad':LandingPad,
         'Block':Block,
         'ReusedCount':ReusedCount,
         'Serial':Serial,
         'Longitude': Longitude,
         'Latitude': Latitude}
```

Create dictionary & dataframe

```
In [31]: # Create a data from launch_dict
         data_falcon9 = pd.DataFrame(launch_dict)
```

Data collection – Web scraping

DataScienceCapstone/Data Collection with Web Scraping Notebook(2).ipynb at n Nirin13/DataScienceCapstone (github.com)

```
In [5]: # use requests.get() method with the provided static_url
        # assign the response to a object
        page = requests.get(static_url)
        page.status_code

Out[5]: 200
```

GET request from Falcon9 launch WIKI page

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(page.text, 'html.parser')
```

Create beautifulsoup object

Extract all column/variable names from the HTML table header

```
headings = []
for key,values in dict(launch_dict).items():
    if key not in headings:
        headings.append(key)
    if values is None:
        del launch_dict[key]

def pad_dict_list(dict_list, padel):
    lmax = 0
    for lname in dict_list.keys():
        lmax = max(lmax, len(dict_list[lname]))
    for lname in dict_list.keys():
        ll = len(dict_list[lname])
        if ll < lmax:
            dict_list[lname] += [padel] * (lmax - ll)
    return dict_list

pad_dict_list(launch_dict,0)
df=pd.DataFrame(launch_dict)
df.head()
```

Create dataframe based on table

Data wrangling

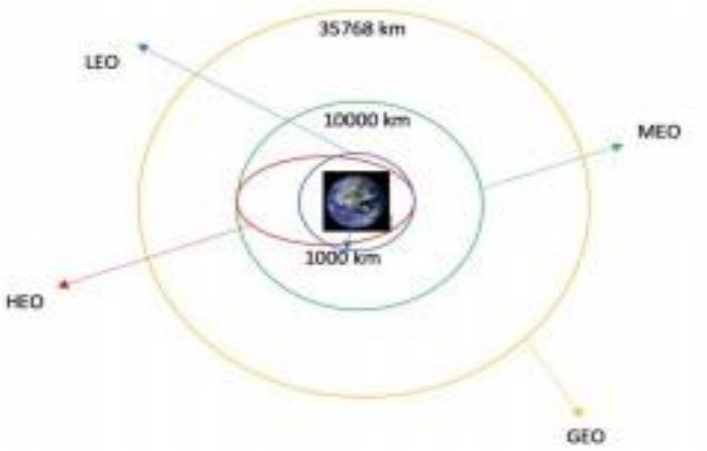
Data analysis: Standardizing, normalizing data, combining data sets

Calculate # of launches at each site

Calclate # and ocurrance of each orbit

Calculate # and occurance of mission outcome per orbit type

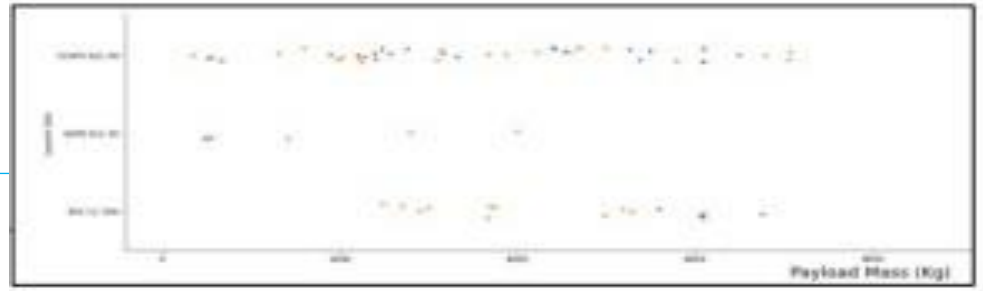
Create new outcome label for data sets

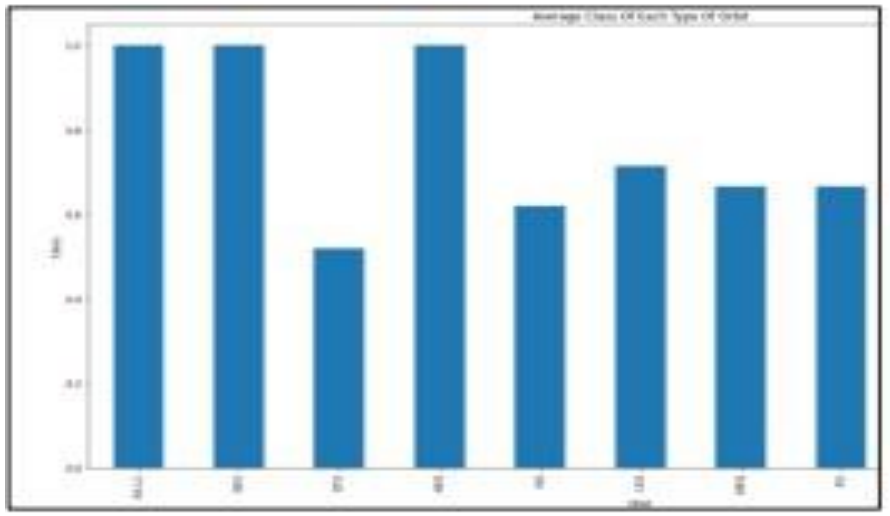


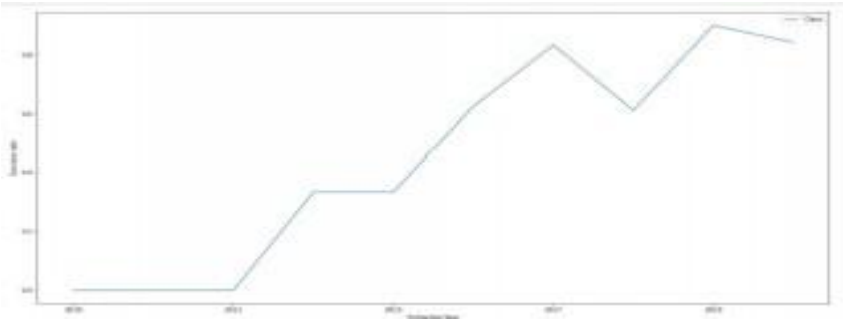
DataScienceCapstone/Data Wrangling EDA Notebook(3).ipynb at main · Nirin13/DataScienceCapstone (github.com)

EDA with data visualization

- Various chart types were plotted to visualize the SpaceX launch data
  - A scatterplot was used to visualize the relationships between flight number and launch site and payload and launch site
  - Scatterplots are useful chart types because it is relatively easy to see relationships and clusterings between variables
  - A bar chart was used to compare the success rate of each orbit type. Bar charts are effective as showing comparisons between variables
  - Lastly, a line chart was used to show average success rate over time. Line charts are effective at presenting performance over time







[DataScienceCapstone/EDA with Data Visualization Notebook(5).ipynb at main · Nirin13/DataScienceCapstone (github.com)]

EDA with SQL

- Various queries were used to obtain information about the data set, including retrieving the following data points
  - Names of unique launch site
  - Total payload mass carried by boosters launched by NASA
  - Date of successful landing outcome in drone ships
  - Names of successful boosters with mass greater than 4000 but less than 600
  - Total number of successful and failed mission outcomes

Build an interactive map with Folium

- To build the folium map, the following map objects were used:
  - The circle object was added to show define the launch site
  - Dataframe launch_outcomes were assigned to colors with green red to show success or failure
  - A line object was added to measures the distance between landmarks

DataScienceCapstone/Visual Analytics with Folium Notebook(6).ipynb at main · Nirin13/DataScienceCapstone (github.com)

Build a Dashboard with Plotly Dash

- Built a dashboard using Flash and Dash web framework
  - Interactions and features include dropdown menu, slider function for ease of data manipulation
- 2 types of graphs were built: pie chart and scatterplot
  - Piechart shows the total launches (and percentage of all launches) by launch site
  - Scatterplot displays the correlative relationship between outcome (success/failure) and payload mass for different booster versions

Predictive analysis (Classification)

Build model:
--Load into a dataset, then visualized the data for easier analysis and review
--Split data into test & training sets
--Choose ML algorithms

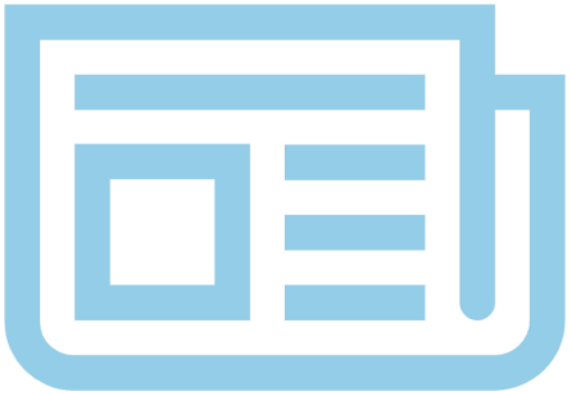
Evaluate model:
--Check model accuracy
--Plot confusion matrix

Imrpove model:
--Tune the algorithms and features

Find the best model
The model with the highest accuracy is selected

DataScienceCapstone/ Machine Learning Prediction Notebook(8).ipynb at main · Nirin13/DataScienceCapstone (github.com)

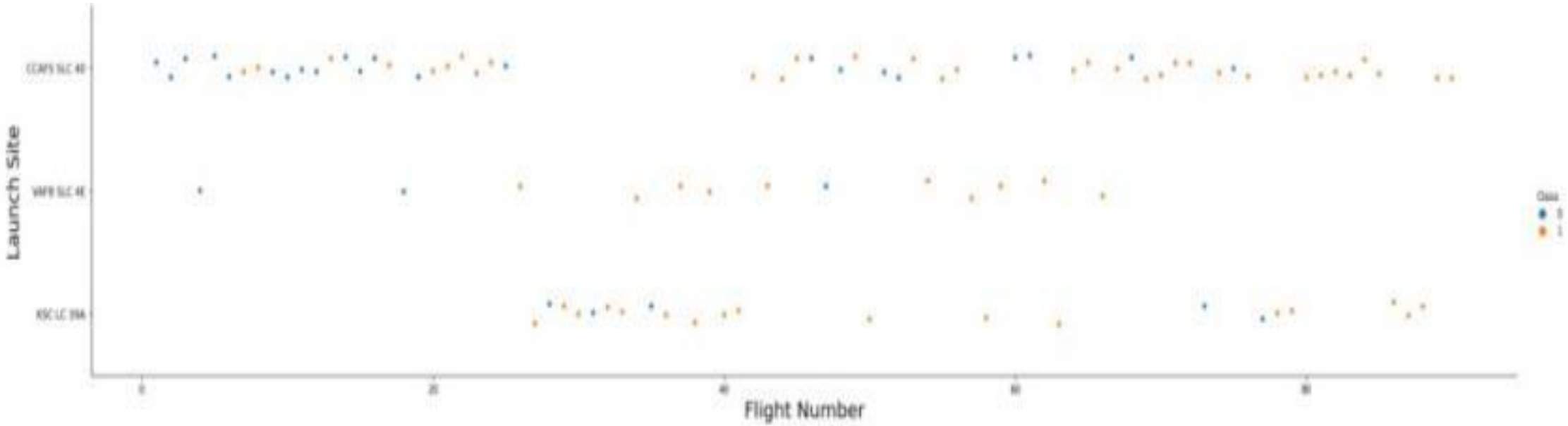
Results



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

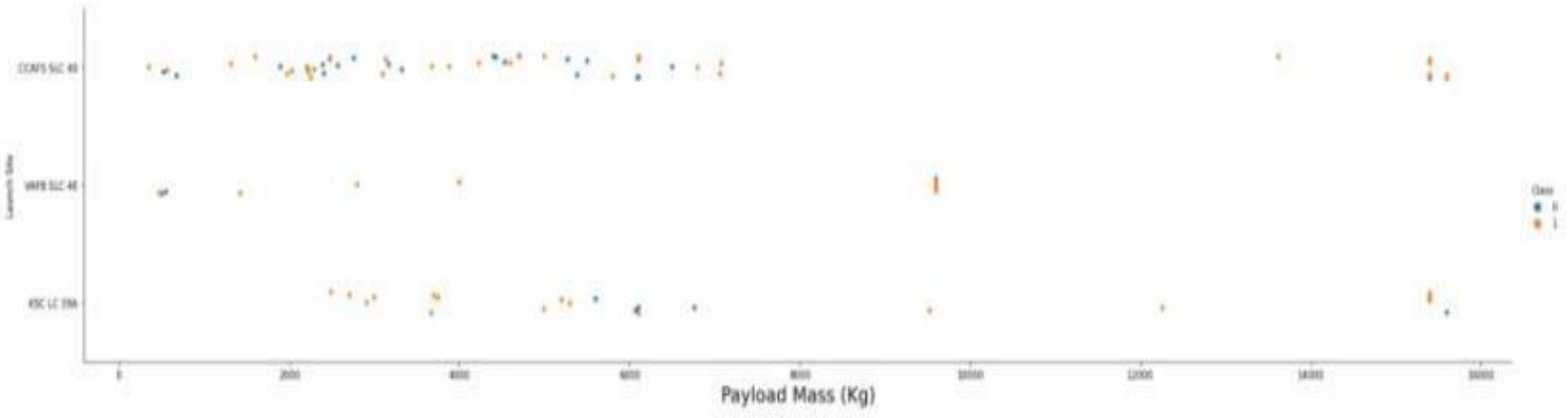
EDA with Visualization

Flight Number vs. Launch Site



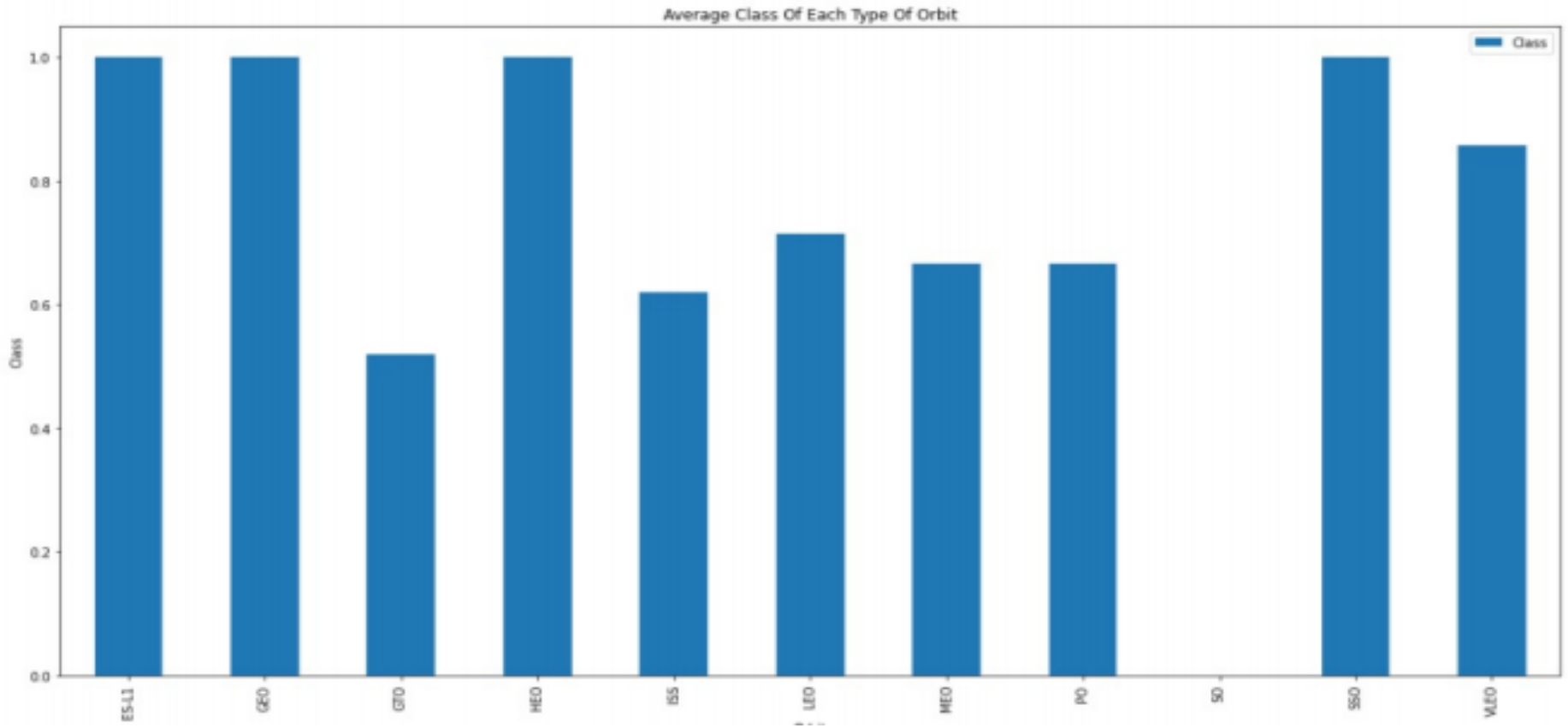
This graph shows the correlations between flight number and the launch site. CCAFS SLC 40 has the most successful launches

Payload vs. Launch Site



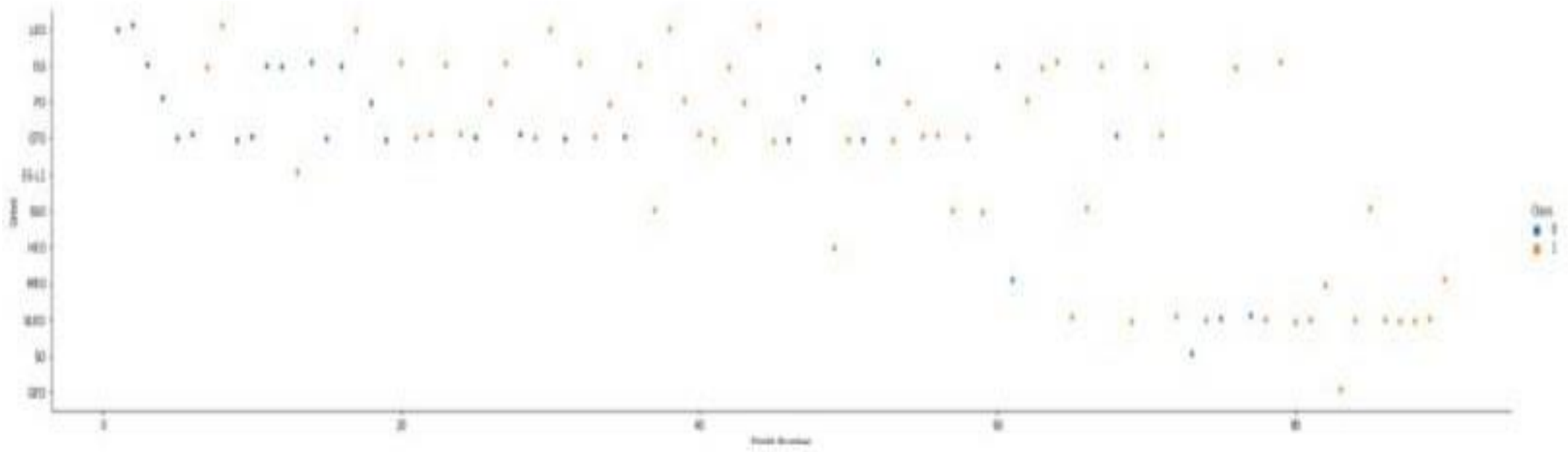
This graph shows the correlations between payload and launch site. There is not enough data to draw a conclusion

Success rate vs. Orbit type



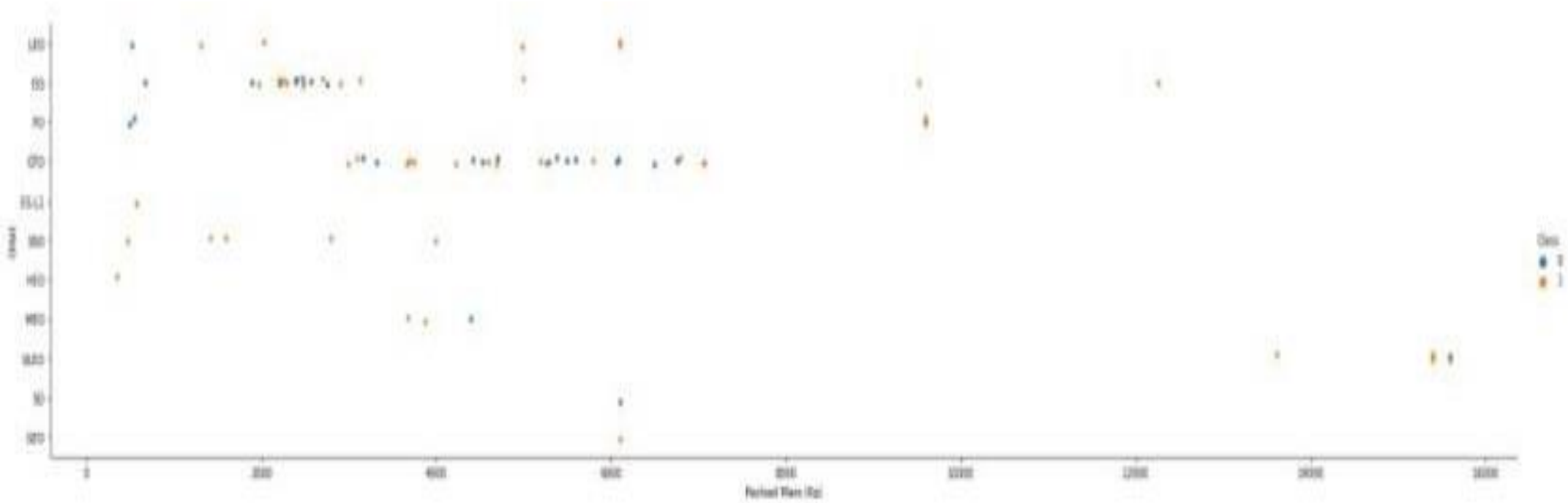
This graph compares the success rate of various orbit types. ESL, GEO, HEO and SSO have the highest success rates

Flight Number vs. Orbit type



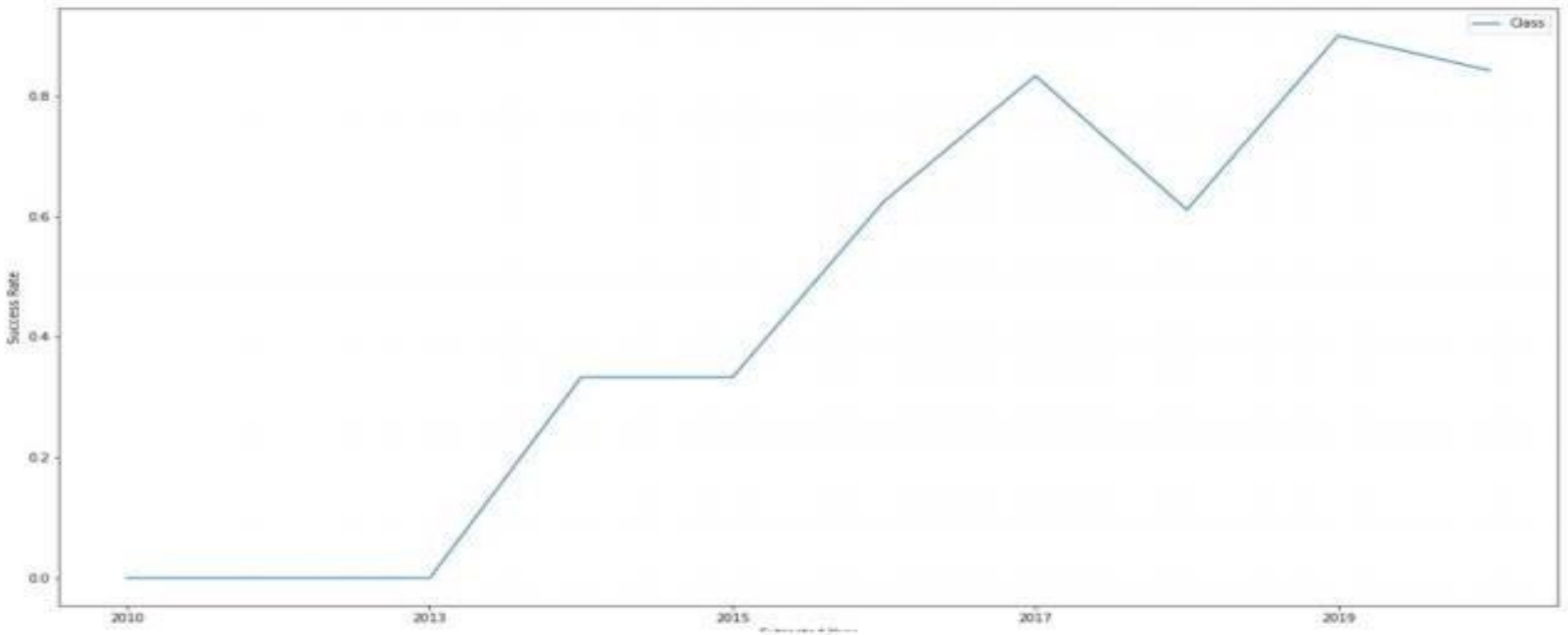
The LEO orbit success is related to the number of flights, but there is little correlation between flight number and GTO orbit

Payload vs. Orbit type



The heavier the payload the more negative the influence on success rates for orbits except GTO and Polar LEO

Launch success yearly trend



As time goes on, the more successful the launches are, wit the exception of a period between 2017 and 2019

EDA with SQL

All launch site names

SQL Query	Results
Select DISTINCT Launch_Site from tblSpaceX	CCAFS LC-40
	CCAFS SLC-40
	CCAFS SLC-40
	KSC LC-39A
	VAFB SLC-4E

This query only returns unique values for launch sites, via ":distinct"

Launch site names begin with `CCA`

SQL Query	Results
%sql SELECT Distinct Launch_Site FROM spacex WHERE Launch_Site LIKE 'CCA%' LIMIT 5	CCAFS LC-40
	CCAFS SLC-40
	CCAFSSLC-40

This query only returns unique values for launch sites, that start with CCA and limits to 5 responses

Total payload mass

SQL Query	Results
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM spacex WHERE CUSTOMER = 'NASA (CRS)'	45596

This query sums the payload mass from spacex table where NASA is the customer

Average payload mass by F9 v1.1

SQL Query	Results
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM spacex WHERE BOOSTER_VERSION = 'F9 v1.1'	2928.4000

This query finds the average payload mass from SpaceX table for the Boostr f9 v1.1

First successful ground landing date

SQL Query	Results
%sql SELECT DATE FROM spacex WHERE Landing_Outcome LIKE '%(ground pad)' LIMIT 1	22-12-2015

This query finds the first successful launch landing

Successful drone ship landing with payload between 4000 and 6000

SQL Query	Results
%sql SELECT Booster_Version FROM spacex WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000) AND Landing_Outcome = 'Success (drone ship)'	F9 FT B1022
	F9 FT B1026
	F9 FT B1021.2
	F9 FT B1031.2

This query finds the list of successful boosters within the specified payload weight range

Total number of successful and failure mission outcomes

SQL Query	Results
%sql SELECT COUNT(Mission_Outcome) AS Success FROM spacex WHERE Mission_Outcome LIKE 'Success%'	100

This query finds the counts the number of successful missions where the outcome contains success

Boosters carried maximum **payload**

SQL Query	Results – booster version	Results -- payload
%sql SELECT Booster_Version,PAYLOAD_MASS__KG_ FROM spacex WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)FROM spacex)	F9 B5 B1048.4	15600
	F9 B5 B1049.4	15600
	F9 B5 B1051.3	15600
	F9 B5 B1056.4	15600
	F9 B5 B1048.5	15600
	F9 B5 B1051.4	15600
	F9 B5 B1049.5	15600
	F9 B5 B1060.2	15600
	F9 B5 B1058.3	15600
	F9 B5 B1051.6	15600
	F9 B5 B1060.3	15600
	F9 B5 B1049.7	15600

This query finds the booster versions that have the maximum payload, using a subquery to find the max payload

2015 launch records

SQL Query	Results – date	Results – landing outomce	Results – booster version	Results – launch site
%%sql SELECT Date, Landing_Outcome, Booster_Version, Launch_Site FROM spacex WHERE Landing_Outcome LIKE 'Failure%' AND DATE LIKE '%%-%%-2015	10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
	14-04-2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

This query combines the landing outcome, booster version and launchsite into a single reuslts where the landing outcome has failed

Rank success count between 2010-06-04 and 2017-03-20

SQL Query	Results
%%sql SELECT COUNT(DATE),DATE FROM spacex WHERE Landing_Outcome LIKE 'Success%' AND (DATE BETWEEN '2010-06-04' AND '2017-03-20') GROUP BY DATE ORDER BY DATE DESC	

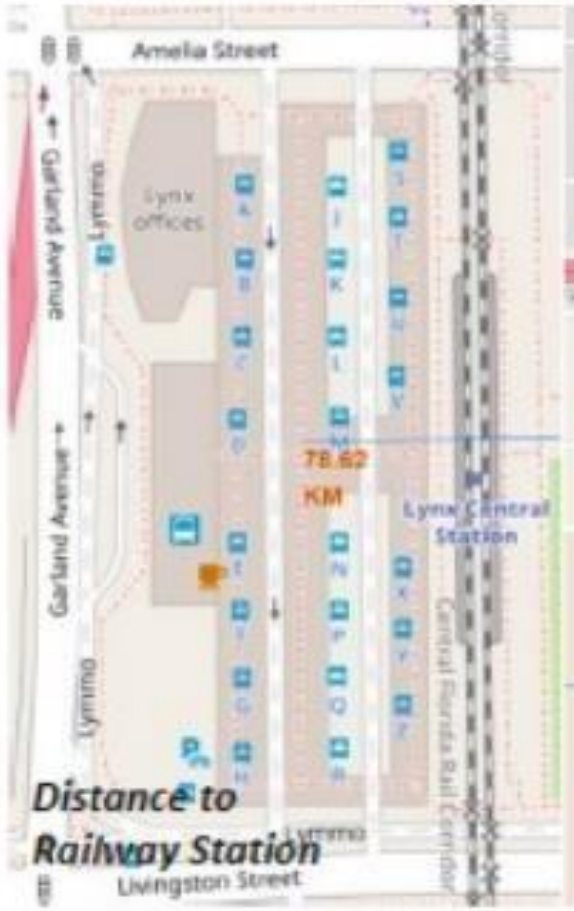
This query did not return any results

Interactive map with Folium

Launch site global map

Successful and failed launches

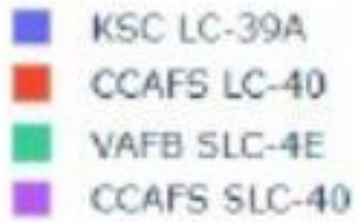
railway



Launch sites are not close to railwats nor highwas. They are however close to coastlines and are a healthy distance away from urban centers

Build a Dashboard with Plotly Dash

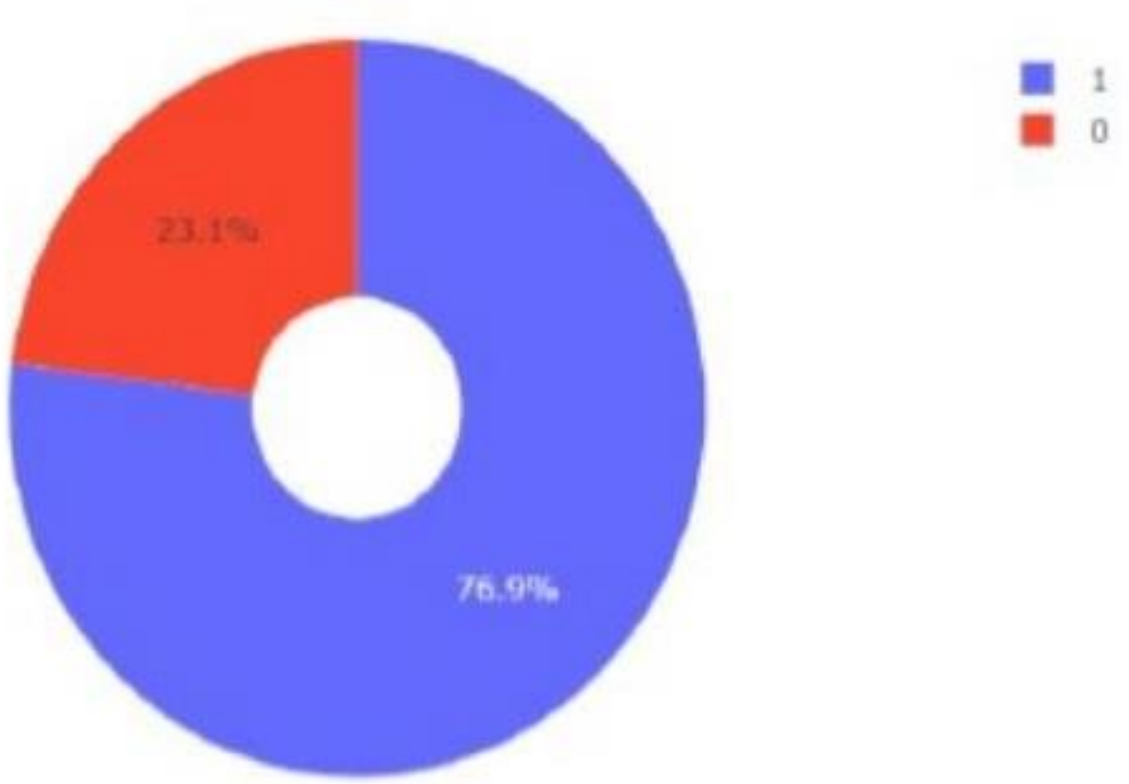
Launch success dashboard



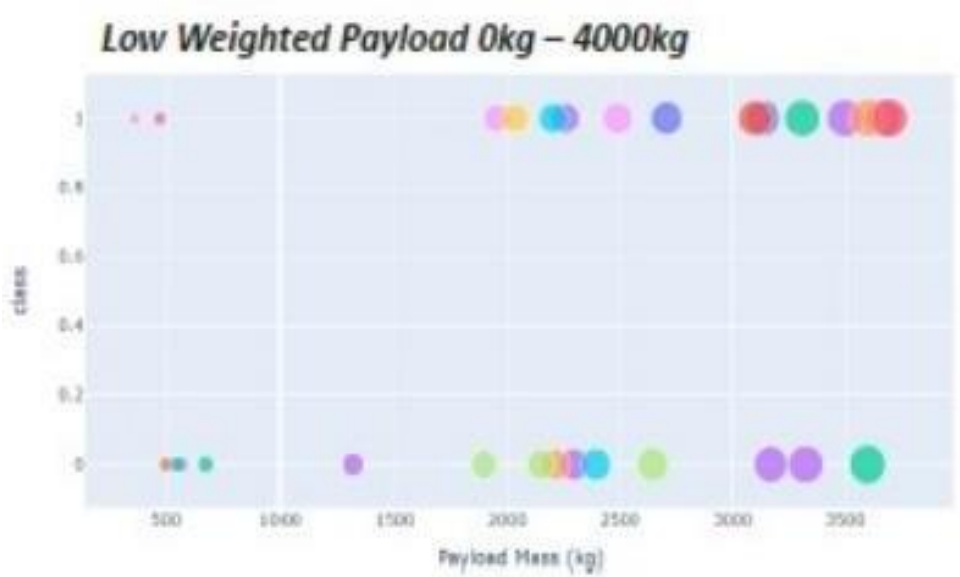
KSC LC-39A had the most successful launch of all launch sites

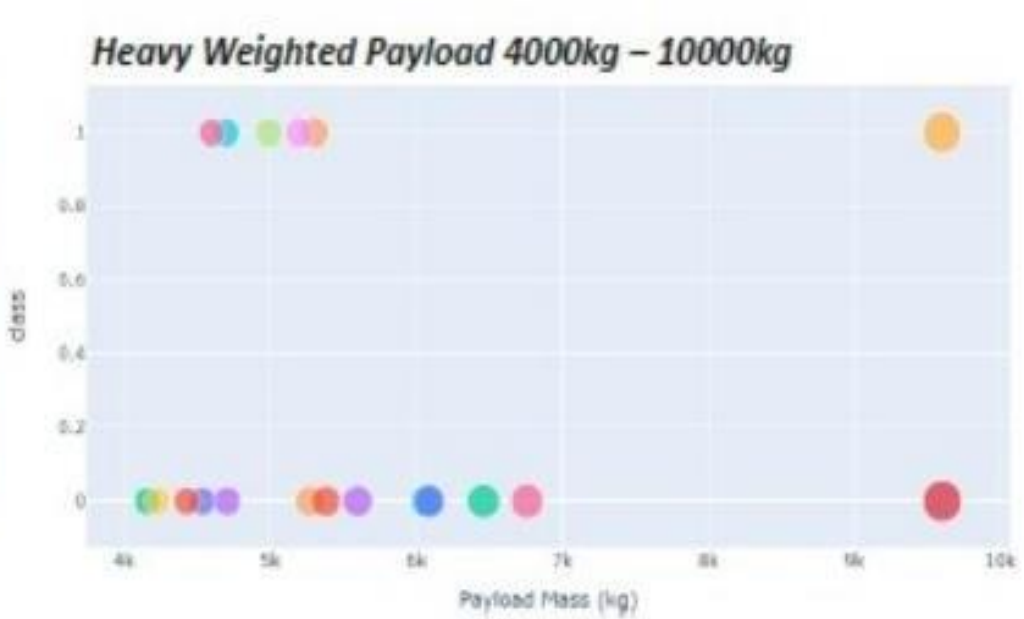
Most successful launch site

KSC LC-39A had the most successful rate of launches



payload





A lower payload correlates to a higher success rate

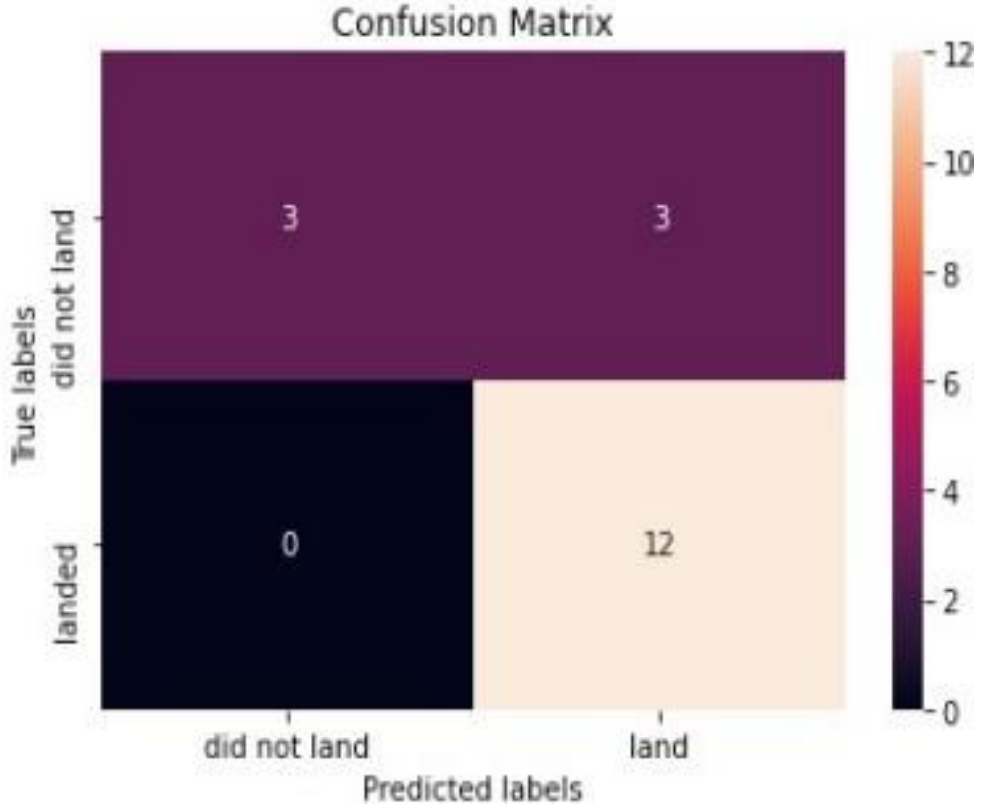
Predictive analysis (Classification)

Classification Accuracy

Decision tree performs the best

Confusion Matrix—Tree

The tree algorithm performs best as it can best distinguish between classes.



CONCLUSION

- For future launches, lighter payloads would more likely be successful when in a GEO, HEO, SSO or ES-L1 Orbi
- SpaceX will be more successful as time goes on
- For this particular data set, the decision tree algorithm was the best suited