



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Diego Emanuel Saucedo Ortega  
01/25/2024



# Outline

---

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

# Executive Summary

---

- Applying data analysis and machine learning, this document presents a detailed SpaceX API inspection for deploying a boosters retrieval success predictor, especially Falcon 9 boosters.
- Describes features with more impact for a successful landing such as: Payload mass, destination orbit and launch site.
- Analyzing launch site locations and their characteristics such as proximities and landing rates.
- Classification models: logistic regression, support vector machines, decision tree classifier and k-nearest neighbor, obtained a 83.3% accuracy using grid search on test data, mostly guessing successful landings.

# Introduction

---

- The new space exploration company: SpaceY, wants to enter the market with the right foot, so is planning to use previous rocket launches data to analyze the relevant features to a successful launch and creating tools for predicting results.
- SpaceX API data is used for analyzing relevant features. This information is recollected with web scrapping and data wrangling with Python library 'pandas'.
- Later, with the clean dataset applying data visualization to appreciate correlation between features, using libraries like folium, pyplot, seaborn and dash.
- Finally, using this data set and relevant features to fit different machine learning models in Python with scikitlearn and determine the best model for predicting successful boosters landing.



Section 1

# Methodology

# Methodology

---

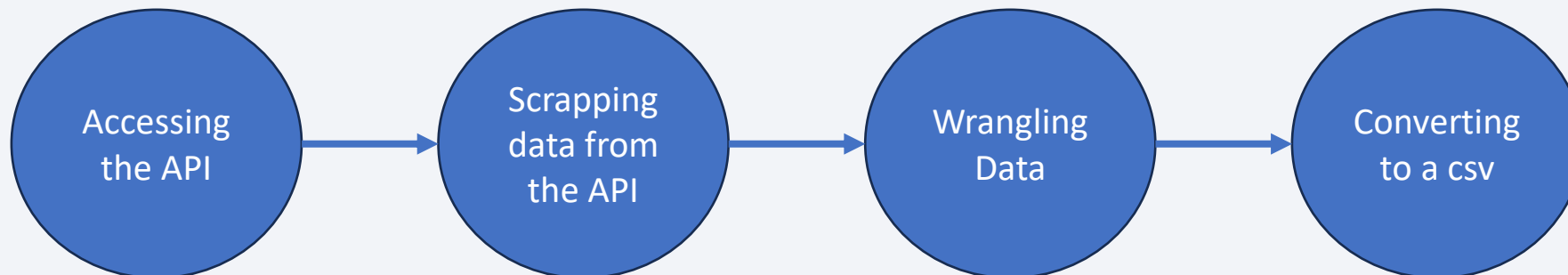
## Executive Summary

- Data collection methodology:
  - Recollected from SpaceX API applying Python requests, then converting the output into a BeautifulSoup to search html tables easily.
- Perform data wrangling
  - Supporting on pandas library, substituting odd values(null,NaN) with mean or frequent value, besides converting categorical values into one bit separated columns.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Splitting clean data into test and train, then implementing a pipeline with Grid Search, Standard Scaler and multiple prediction models with different parameters.

# Data Collection

---

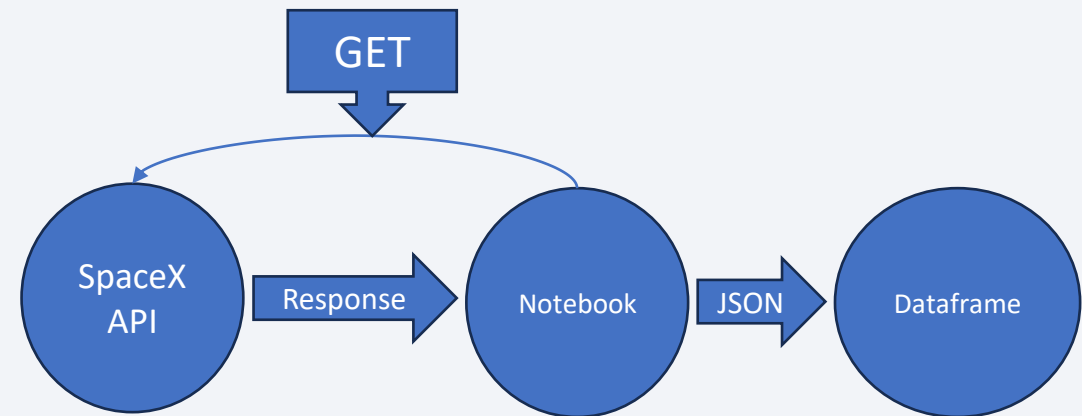
- The data collection begins with the SpaceX API that includes information about previous launches such as: names, locations, dates, boosters characteristics and mission results.
- For extracting data from the API is required to use web scrapping. Libraries like bs4, html5lib and requests in Python simplifies the process of copy-pasting tables into a dataset.
- Dataset are wrangled using pandas. Wrangling data consists in change invalid inputs in the dataset with values that won't affect results, selecting the necessary features for the predictor and converting the dataset to a csv file.



# Data Collection – SpaceX API

---

- First, accessing to SpaceX API with this url:  
<https://api.spacexdata.com/v4/launches/past>
- With Python requests, use the GET protocol to access the content in the API.
- Turn it into a dataframe with `.json_normalize()`.



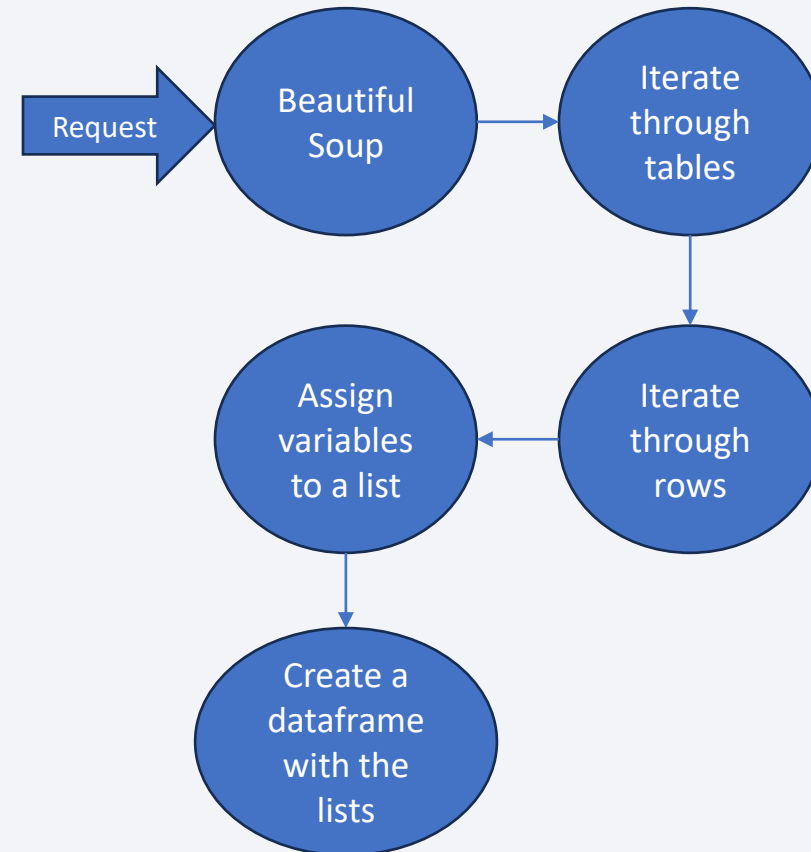
[Github/Data Collection](#)



# Data Collection - Scraping

---

- To reach the desired data is necessary to identify the tables in the HTML, for this we use a BeautifulSoup element that constructs a net with nested HTML tag so is easier to reach those tables.
- Use the `find_all()` method of the soup to get all the `<table>` tags in the HTML.
- Add each value of the row to a specific list.
- Include all the lists to a pandas dataframe.



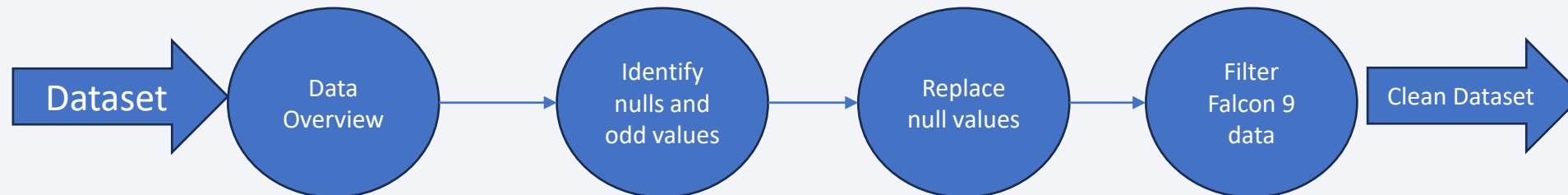
# Data Wrangling

---

- Starting with an overview of the categorical values in features: Landing Sites, Landing Outcomes and Orbit.
- Use `.isnull().sum()` to identify null values in the dataset. In this case, the features Payload Mass and Landing Pad. The null values in Payload Mass are replaced by the mean. Landing Pad null values stay.
- Finally, 'Falcon 9' data is selected with the condition:

Booster Version == 'Falcon 9'

[Github/ Data Wrangling](#)



# EDA with Data Visualization

---

For a useful exploratory data analysis, a visualization offers better insights about the behavior of data.

Features are analyzed with:

- Scatter/Regression plot: to visualize correlation between two features.
- Bar chart: to understand the success rate of launches according to the destination orbit.
- Linear plot: to watch the growth of success rate through the years
- Map: to situate Launch sites and their launches class.

[Github/ EDA Data Visualization](#)

# EDA with SQL

---


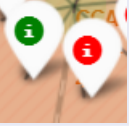
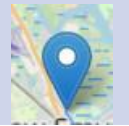

For exploring data, the next queries were executed:

- Unique launch sites names.
- Launch records from sites beginning in 'CCA'
- Total Payload Mass launched by NASA
- Average Payload Mass for Falcon9 v1.1
- The date of the first success in ground pad landing
- List of boosters with success in drone ship landing with a Payload Mass between 4000 and 6000
- Number of failed and success missions
- Name of the boosters that carried the maximum Payload Mass
- Month and Booster version of drone ship failed landing in 2015
- Ranking of counts of landing outcomes

# Build an Interactive Map with Folium

---

Maps presents the launch sites locations, the success and failed launches and proximities.

Icon	Name	Description
	Launch Site	Indicates a location where launches are realized.
	Launch Success/Fail	Indicates with green if a launch is successful and red if it's a failure.
	Interest Location	Indicates either a railway, highway or city.
	Distance line	Trace the distance from a launch site to a interest location.

[Github/ Folium maps](#)



# Build a Dashboard with Plotly Dash

---

Including a dashboard with two plots:

- Pie plot: Indicates success launches for all sites, additionally successful and failed launches for each site.
- Scatter plot: Visualize the different Payload Mass normally launched in every launch site.

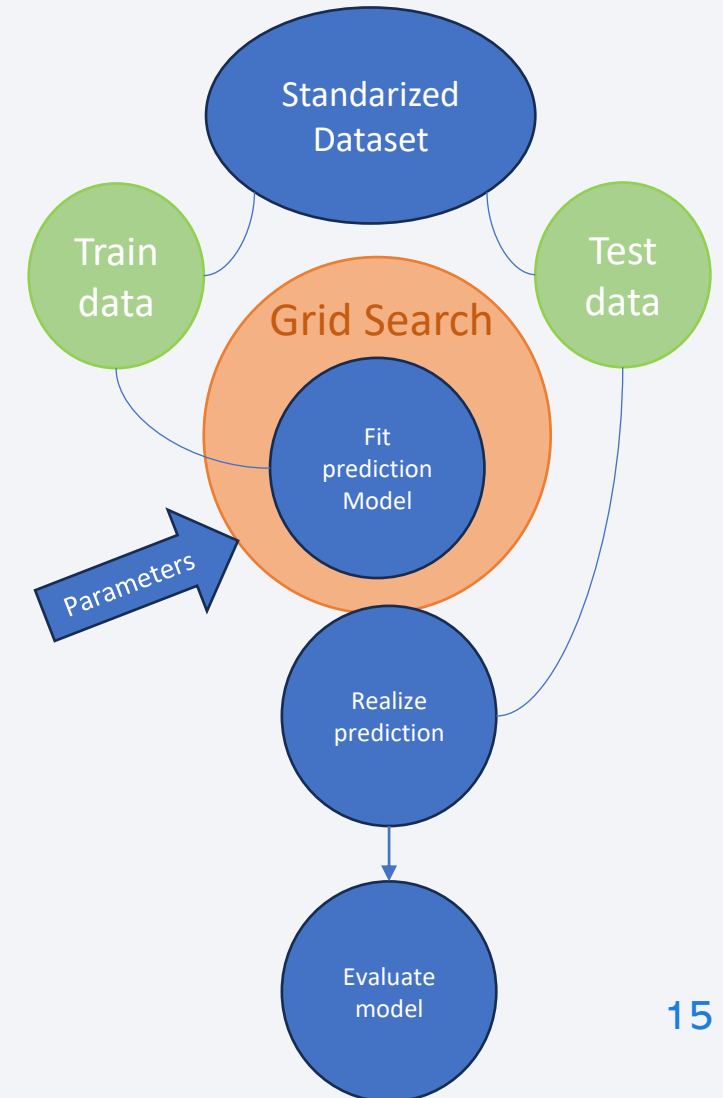
In the case of the pie plot, there is a dropdown to select the site of the desired info. For the scatter, besides the dropdown, is necessary to use the range slider for selecting the range of the payload.

[Github/ Dashboard](#)

# Predictive Analysis (Classification)

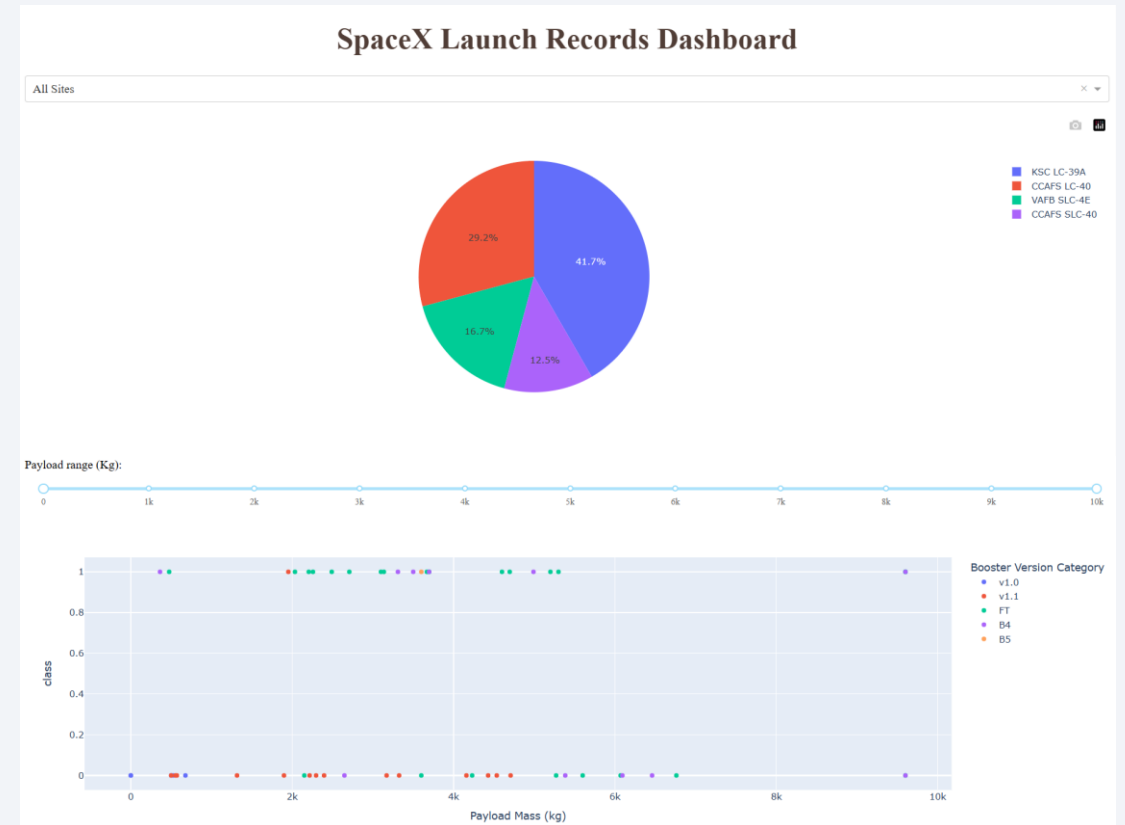
- First step is splitting data into test and train considering to drop the target from the features dataset.
- Standardize the values in the dataset with StandardScaler and fit transform.
- Deploy a pipeline with GridSearchCV to determine the best parameters for the expected prediction model.
- Models to select are: Logistic Regression, Classification Tree, K-Nearest Neighborhood, Support Vector Machine.
- Compare the scores and confusion matrix of every model to select the one that fits the data better.

[Github/ Machine Learning](#)



# Results

- The exploratory data analysis indicates that launch site, destination orbit and payload mass are features with great importance in the booster landing result.
- Predicting models with the best parameters given by the grid search have an accuracy score of 83% with majority of prediction errors are false positives.





The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

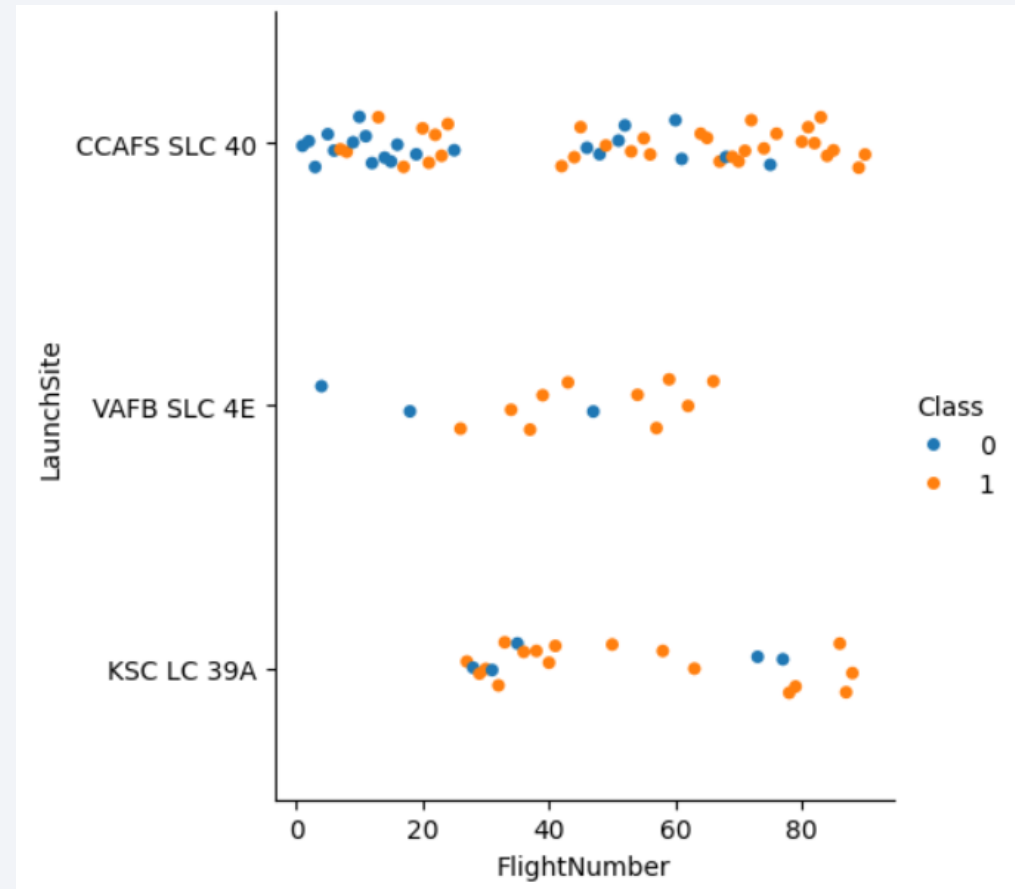
Section 2

# Insights drawn from EDA



# Flight Number vs. Launch Site

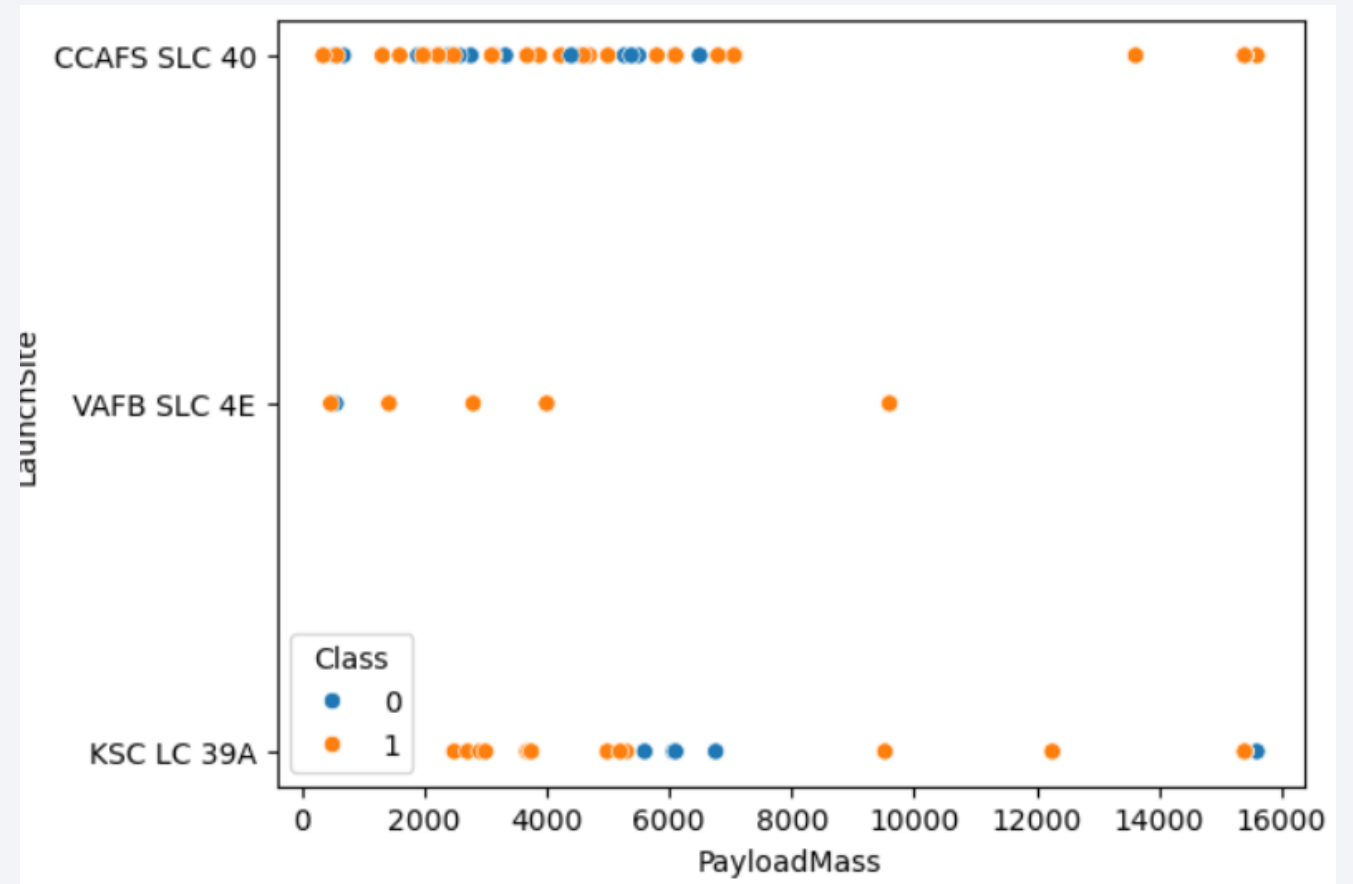
- In this graph, is appreciable that KSC-LC-39A location has more succeed launches than failed.
- CCAFS-SLC-40 has a great number of launches in comparison of the other two locations.





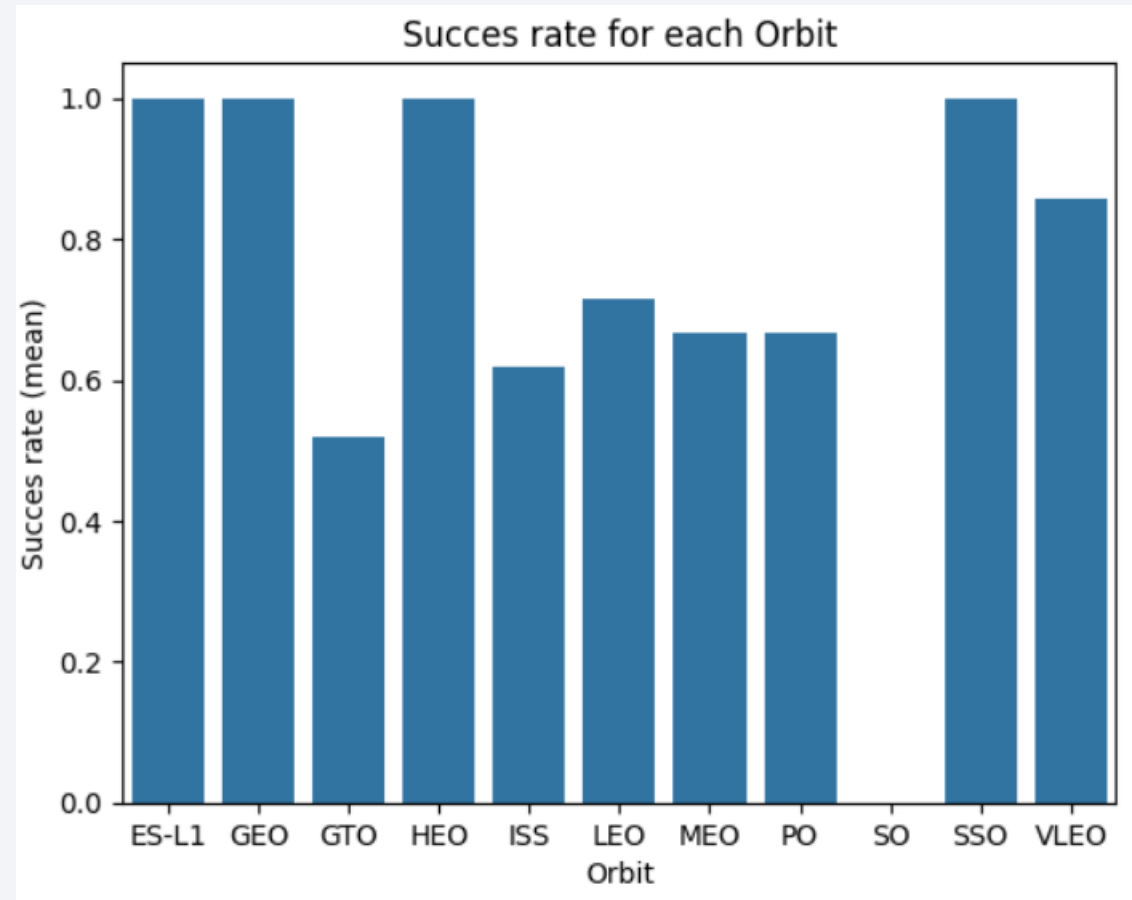
# Payload vs. Launch Site

- CCAFS-SLC-40 and KSC-LC-39A were used for heavy payload mass launches.
- VAFB-SLC-4E has success in short payload mass launches.
- CCAFS-SLC-40 hosted a wide range of payload mass launches.



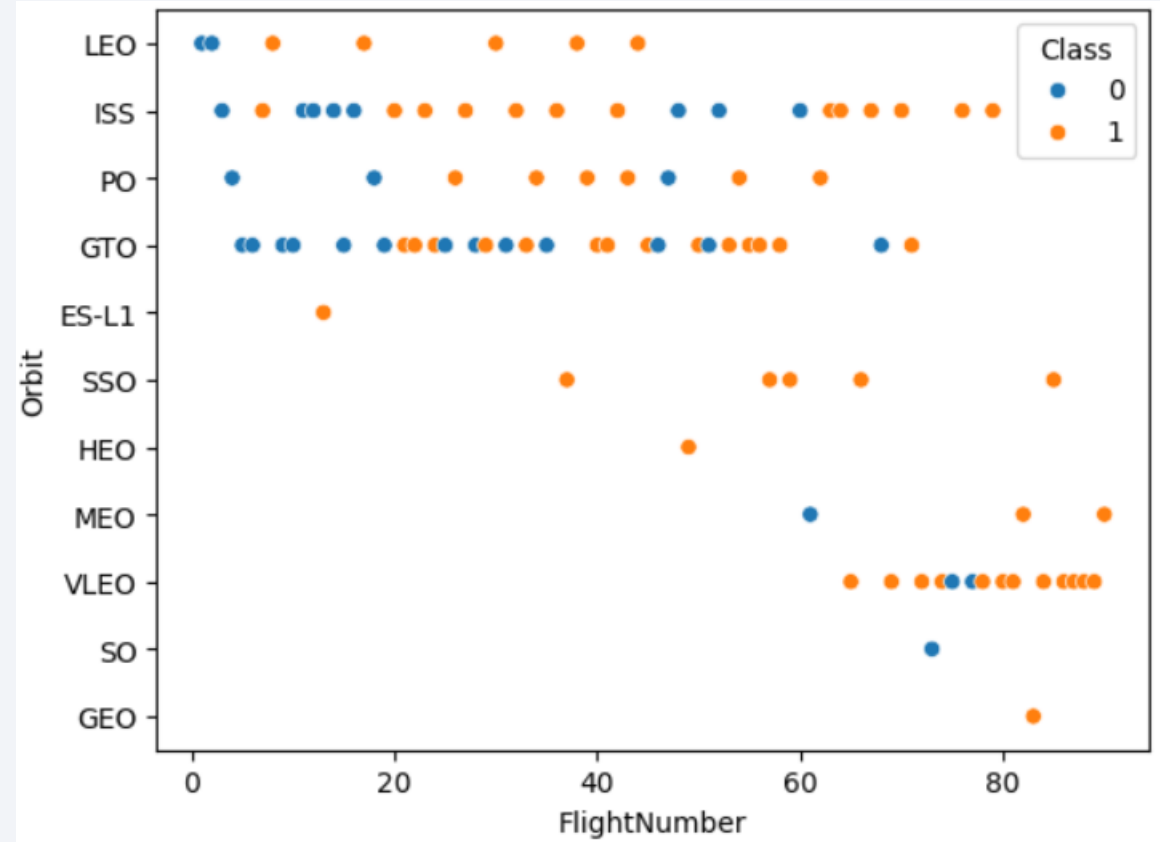
# Success Rate vs. Orbit Type

- In this chart we can appreciate orbits with perfect rate (1.0), ES-L1, GEO, HEO and SSO.
- The orbit with worst success rate is SO.



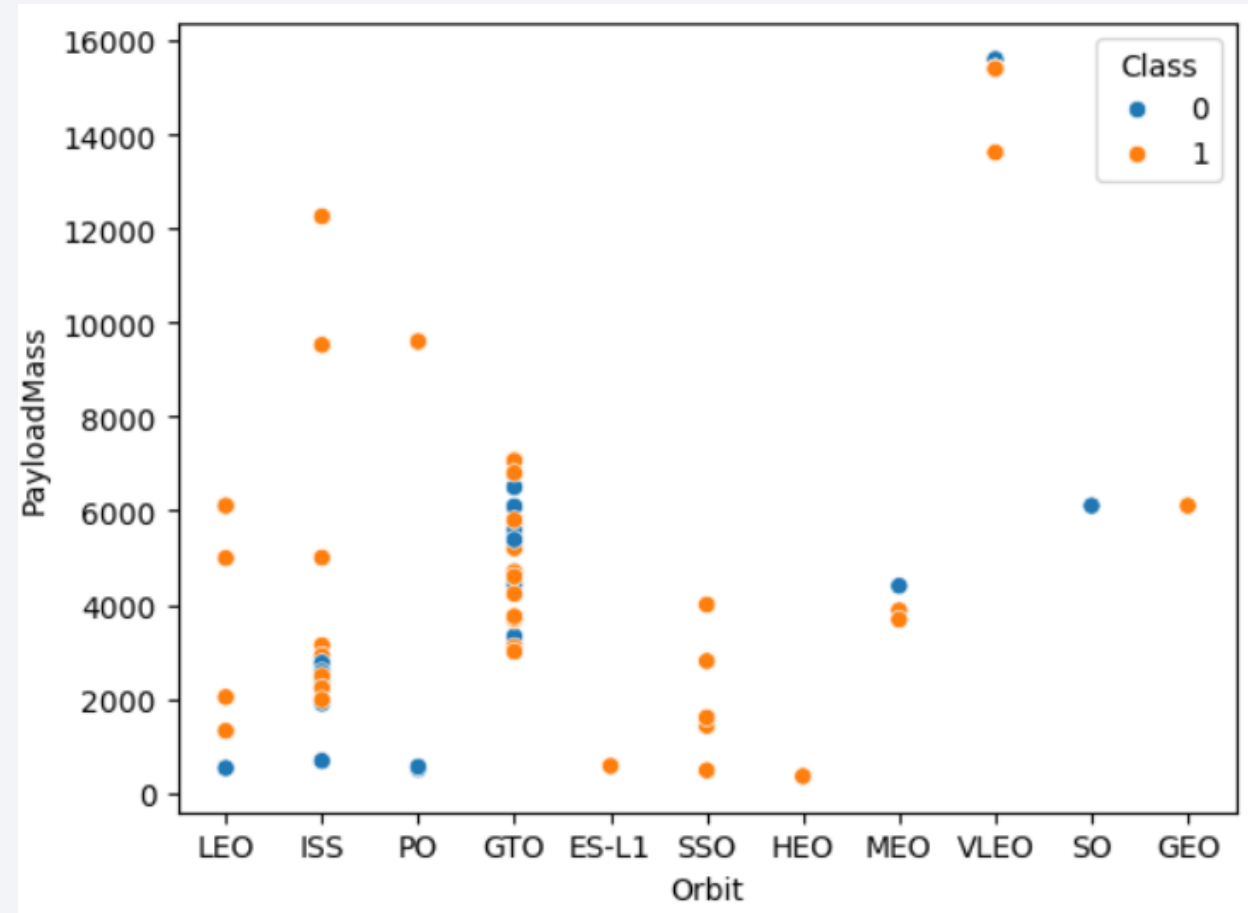
# Flight Number vs. Orbit Type

- Is visible that ES-L1 and SO have one launch, so their success rate is not reliable.
- LEO and VLEO have a great quantity of succeed launches.
- GTO, ISS and VLEO are the most selected destination orbits.



# Payload vs. Orbit Type

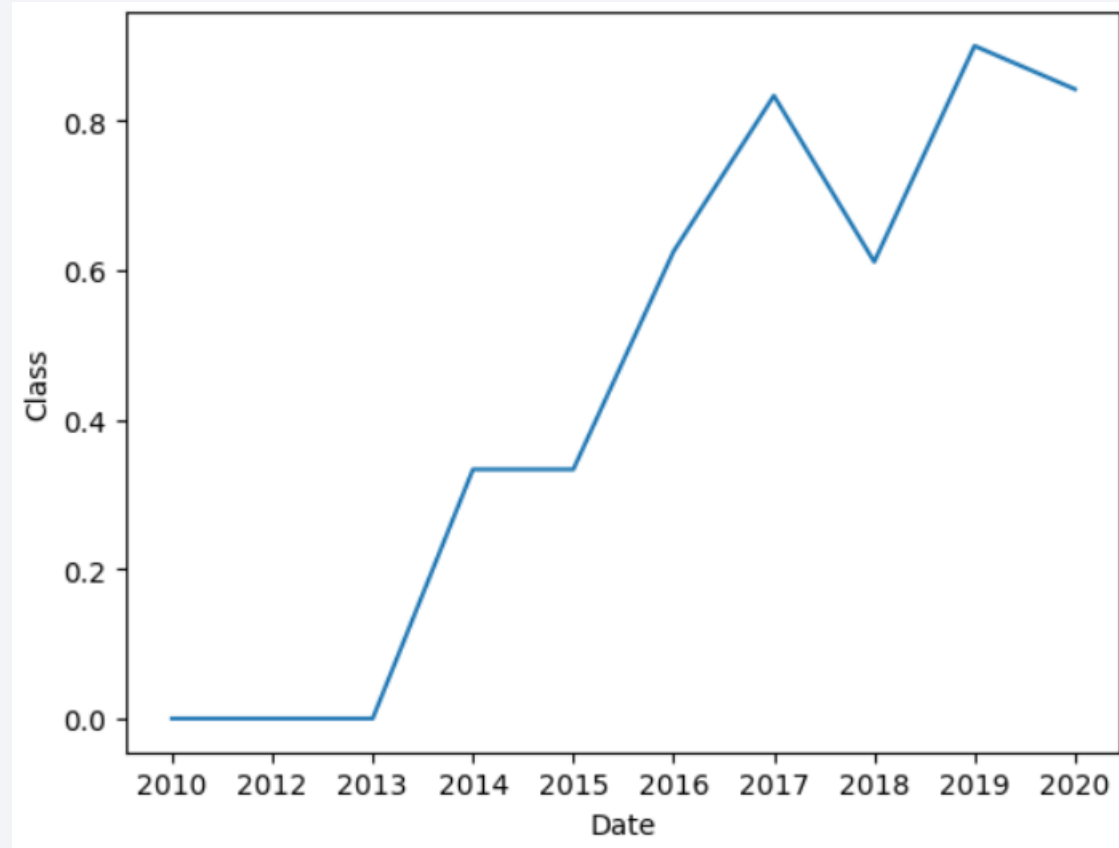
- VLEO and ISS are the destination orbits with heavy payload mass.
- GTO launch payload mass is currently between 8000 and 3000.



# Launch Success Yearly Trend

---

- Its appreciable a positive growth in success rate through the years.
- The fall in 2019 could be a consequence of pandemic.





# All Launch Site Names

---

- There are four launch sites in the dataset. DISTINCT returns the different values selected, in this case the launch site.

## Task 1

Display the names of the unique launch sites in the space mission

```
] : %sql select DISTINCT(Launch_Site) FROM SPACEXTBL
```

```
* 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'

- There are two launch sites with that begin with 'CCA': 'CCAFS LC-40' and 'CCAFS SLC-40'. The result is limited to 5 rows.

```
In [8]: %sql select * from SPACEXTBL where Launch_Site LIKE 'CCA%' LIMIT 5
```

\* sqlite:///my\_data1.db  
Done.

Out[8]:

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

# Total Payload Mass

---

- Total Payload Mass launched on Falcon 9 boosters by NASA is equal to 45,596 kg.

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select SUM(PAYLOAD_MASS_KG_) FROM SPACEXTBL where Customer = 'NASA (CRS)'
```

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

```
Done.
```

```
SUM(PAYLOAD_MASS_KG_)
```

```
45596
```

# Average Payload Mass by F9 v1.1

---

- Taking the average payload mass of F9 v1.1 results 2534.66 kg, so F9 v1.1 launches were with low payload mass.

Display average payload mass carried by booster version F9 v1.1

```
: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

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

```
Done.
```

```
: avg(PAYLOAD_MASS__KG_)
```

```
2534.6666666666665
```

# First Successful Ground Landing Date

---

- MIN returns the lowest value of the column, in this case the date. Is necessary to filter the selection to landing outcomes equal to 'Success (ground pad)'.

```
%sql select min(Date) from SPACEXTBL 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

---

- This is the result of boosters with a medium-lower payload mass that successfully landed on a drone ship.

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

```
%%sql select Booster_Version from SPACEXTBL
where Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

```
Booster_Version
```

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

# Total Number of Successful and Failure Mission Outcomes

---

- In 101 mission outcomes, there are 100 success (one reported as 'payload status unclear') and one in flight failure. GROUPBY is used to assign the value to the count of that value.

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome,COUNT(*) from SPACEXTBL 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

- This is the result of two queries:
- The subquery that obtains the MAX payload mass in the data set.
- The query that select the booster's name with a payload mass equal to the subquery.

List the names of the booster\_versions which have carried the maximum payload mass. Use a subquery

```
%%sql select Booster_Version from SPACEXTBL
where PAYLOAD_MASS_KG_ =
(select MAX(PAYLOAD_MASS_KG_) from SPACEXTBL)
```

```
* 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

---

- Selecting the month (in date, the sixth position, the next two characters), booster version, launch site and landing outcome, the last one with value equal to 'Failure (drone ship)'. It's important to filter this selection by comparing the last four characters of Date with '2015'.

List the records which will display the month names, failure landing\_outcomes in drone ship, booster versions, launch\_site for the months in year 2015.

**Note:** SQLite does not support monthnames. So you need to use `substr(Date, 6,2)` as month to get the months and `substr(Date,0,5)='2015'` for year.

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

\* sqlite:///my\_data1.db

Done.

	substr(Date,6,2)	Booster_Version	Launch_Site	Landing_Outcome
	01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
	04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

---

- This is the count of every landing outcome between 2010-06-04 and 2017-03-20 ordered from less to most frequent.

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(*) from SPACEXTBL
where Date between '2010-06-04' and '2017-03-20' GROUP BY Landing_Outcome
Order by COUNT(*)
```

\* sqlite:///my\_data1.db  
Done.

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

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

# Launch Sites Proximities Analysis



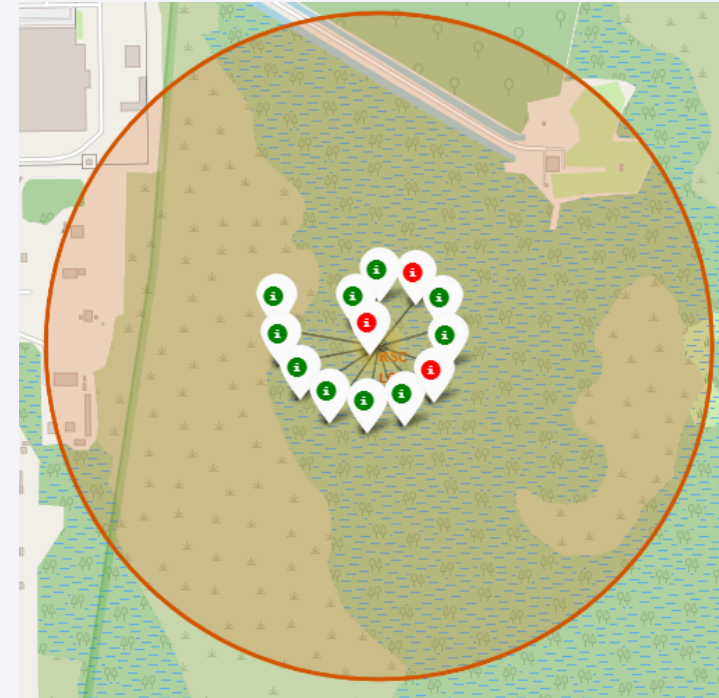
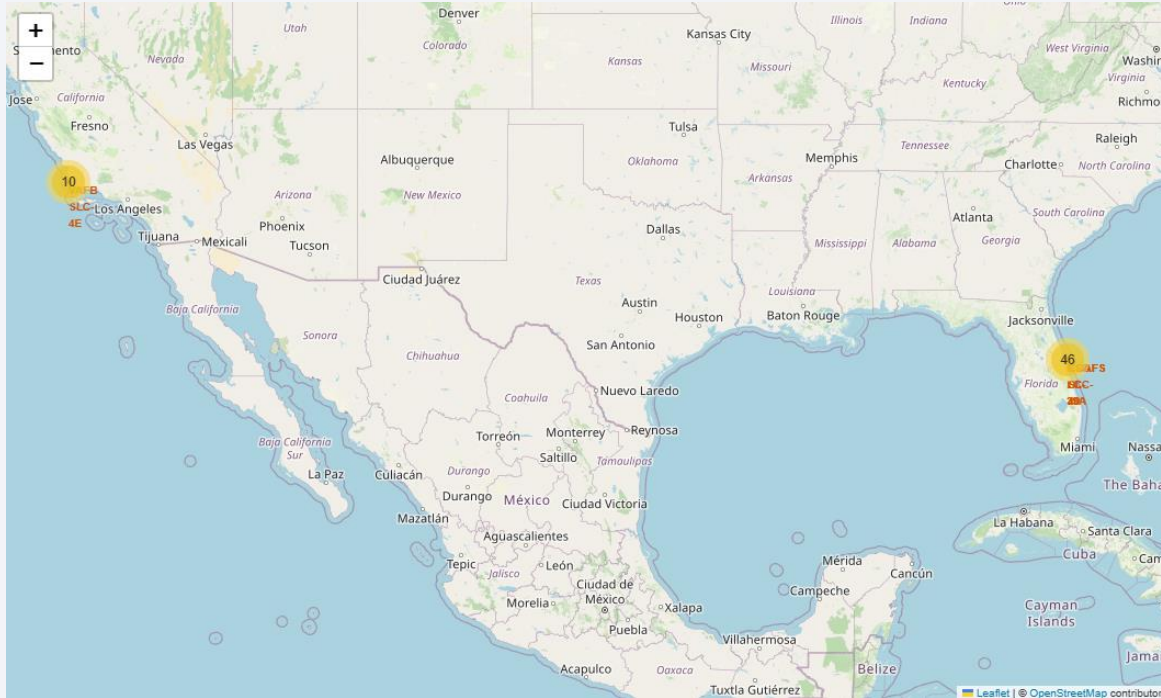
# Launch Site Locations map



It's appreciable that all the launch sites are situated closer to the coasts.

CCAFS launch sites are close to each other.

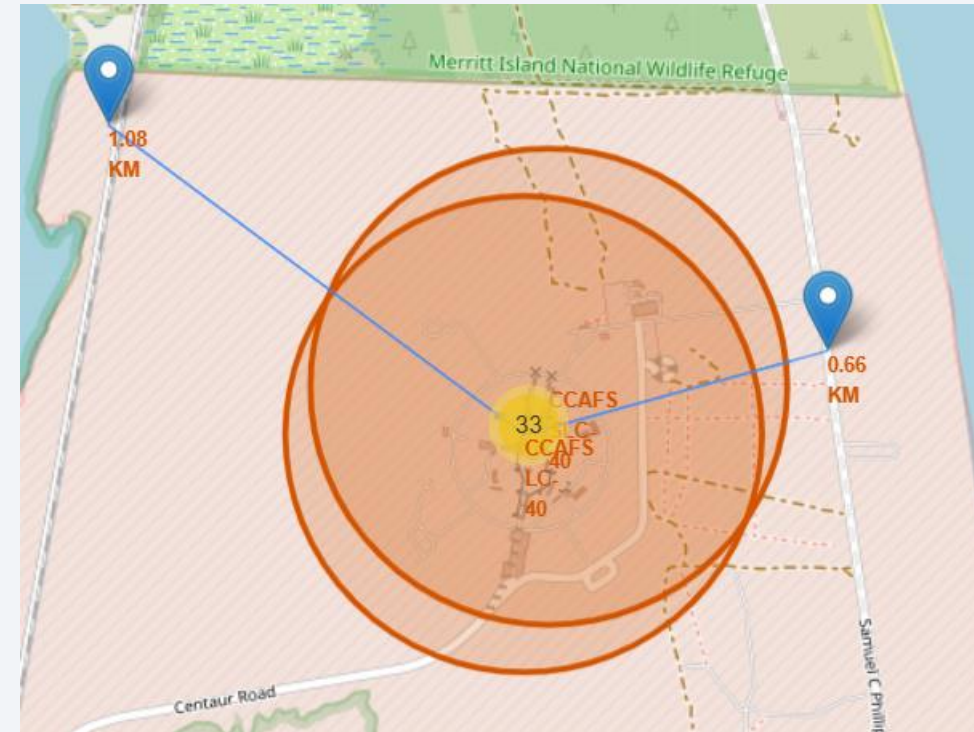
# Successful/Failed Launches map



- It shows a total of 56 launches and their results. 10 on the west coast and 46 on the east.
- The second image shows successful launches in green and unsuccessful launches in red. (Launches shown from KSC LC-39A)



# Launch site distance to proximities



- CCAFS launch sites have at least a 0.6 km radius away from an entrance. The closest city is 62.82 km away. This, to create a safe zone in case of failed launches.

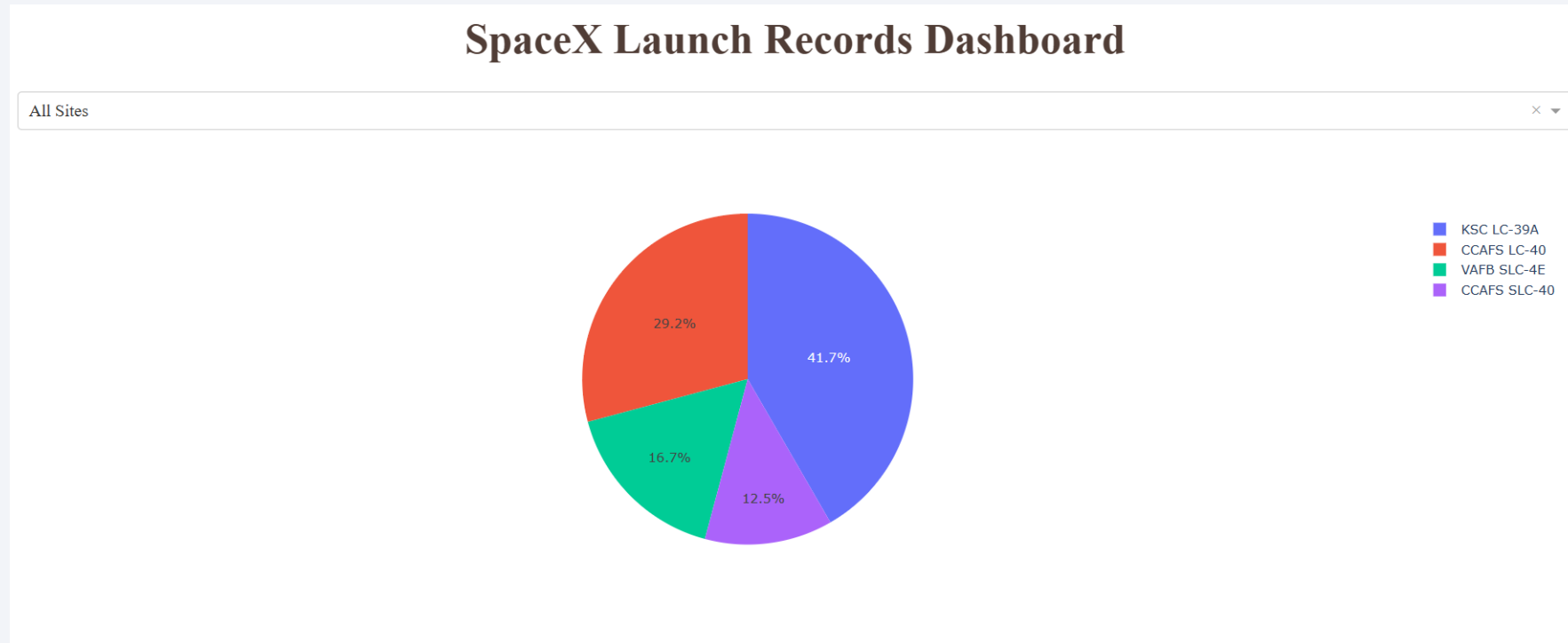


Section 4

# Build a Dashboard with Plotly Dash

# Success proportion of all sites

---



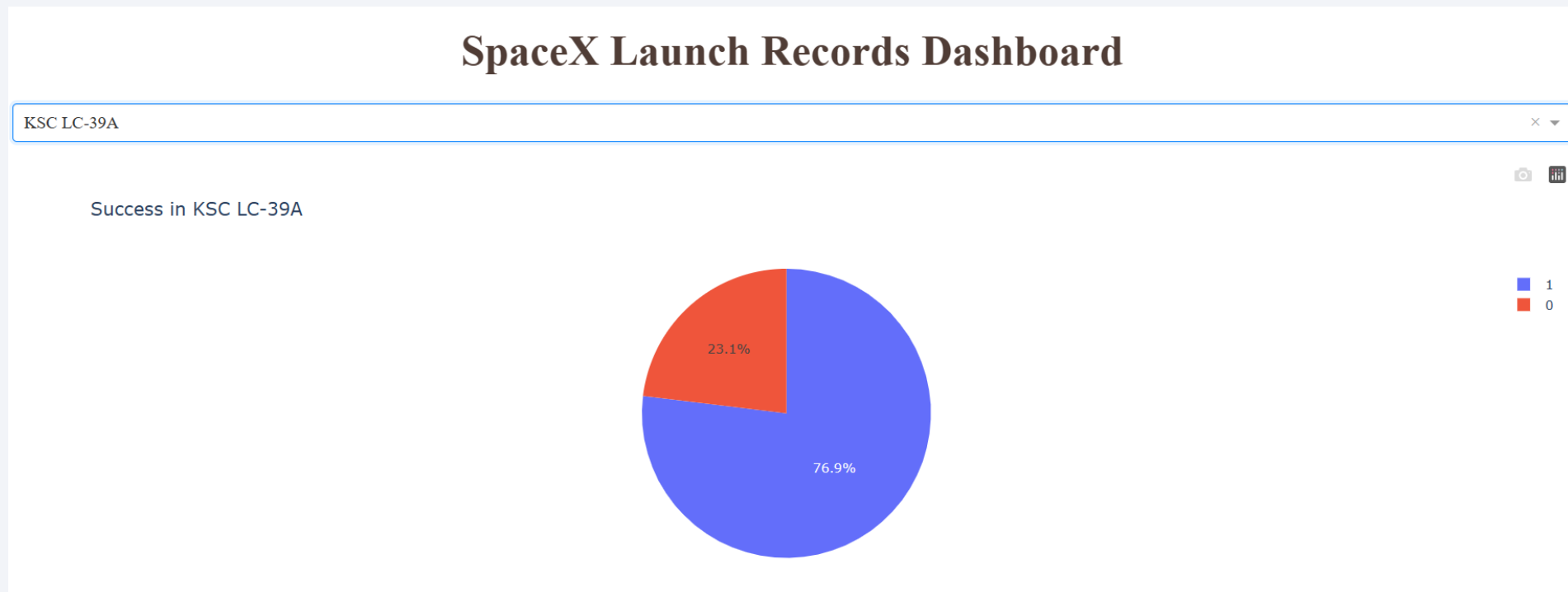
- KSC LC 39A is the launch site with more success proportion.
- CCAFS SLC-40 is the launch site with less success proportion.



# KSC LC 39A success/fail proportions

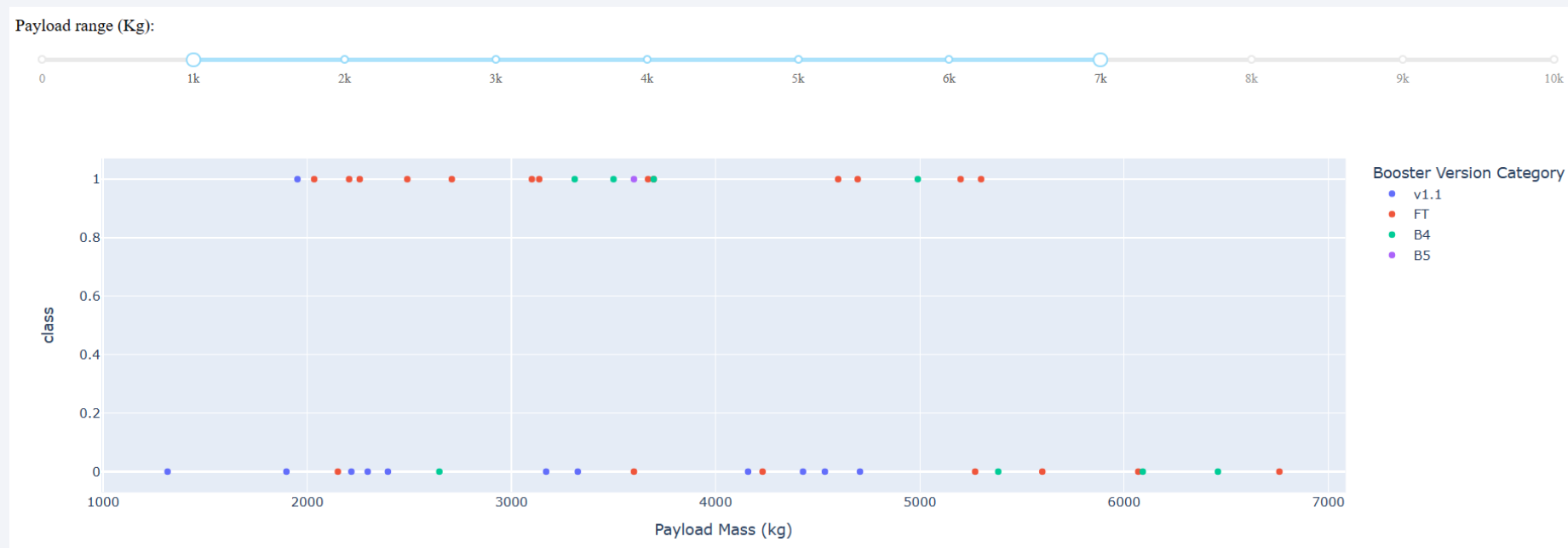
---

- Of the total of launches in KSC LC 39A, 76.9% correspond to success and 23.1% to a fail.



# Payload Mass related to success and fails.

- The Payload Mass range selected is the one with more launches and it's appreciable more failures than success.
- FT and v1.1 sum the majority of launches in this payload range.



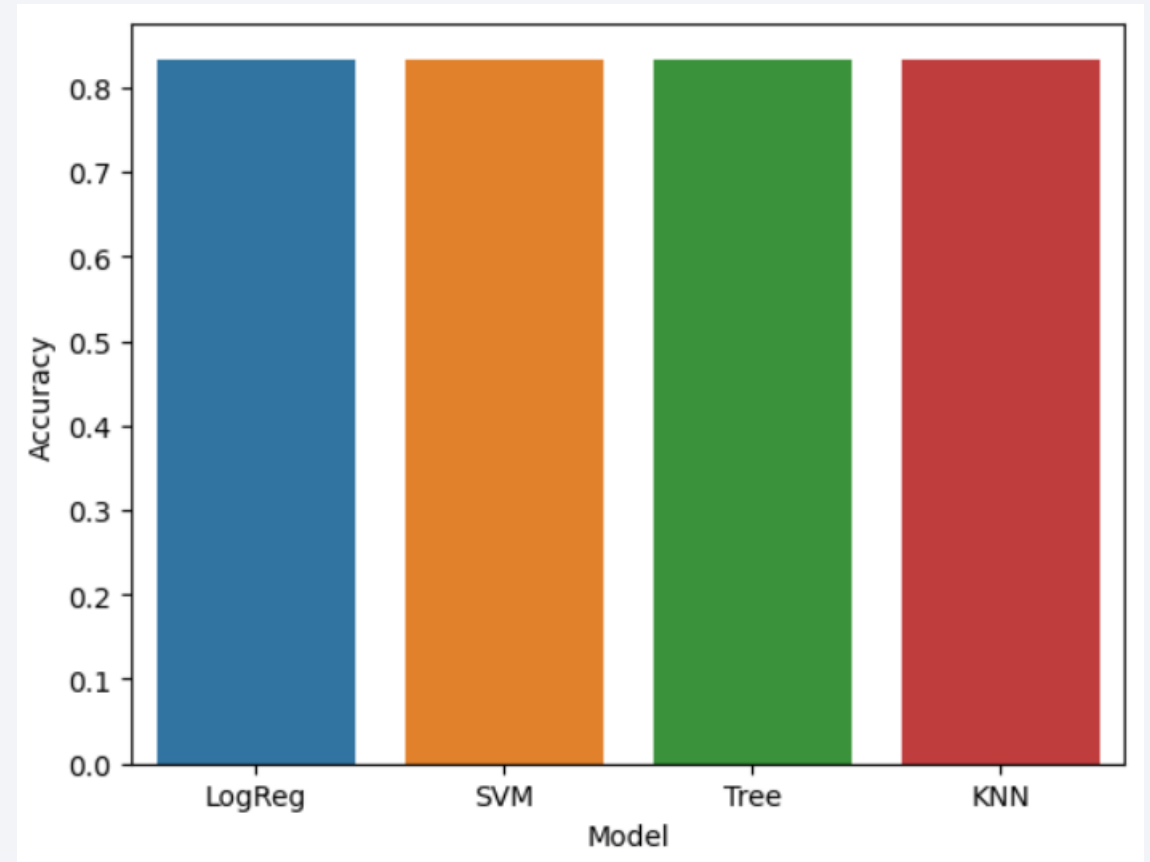
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

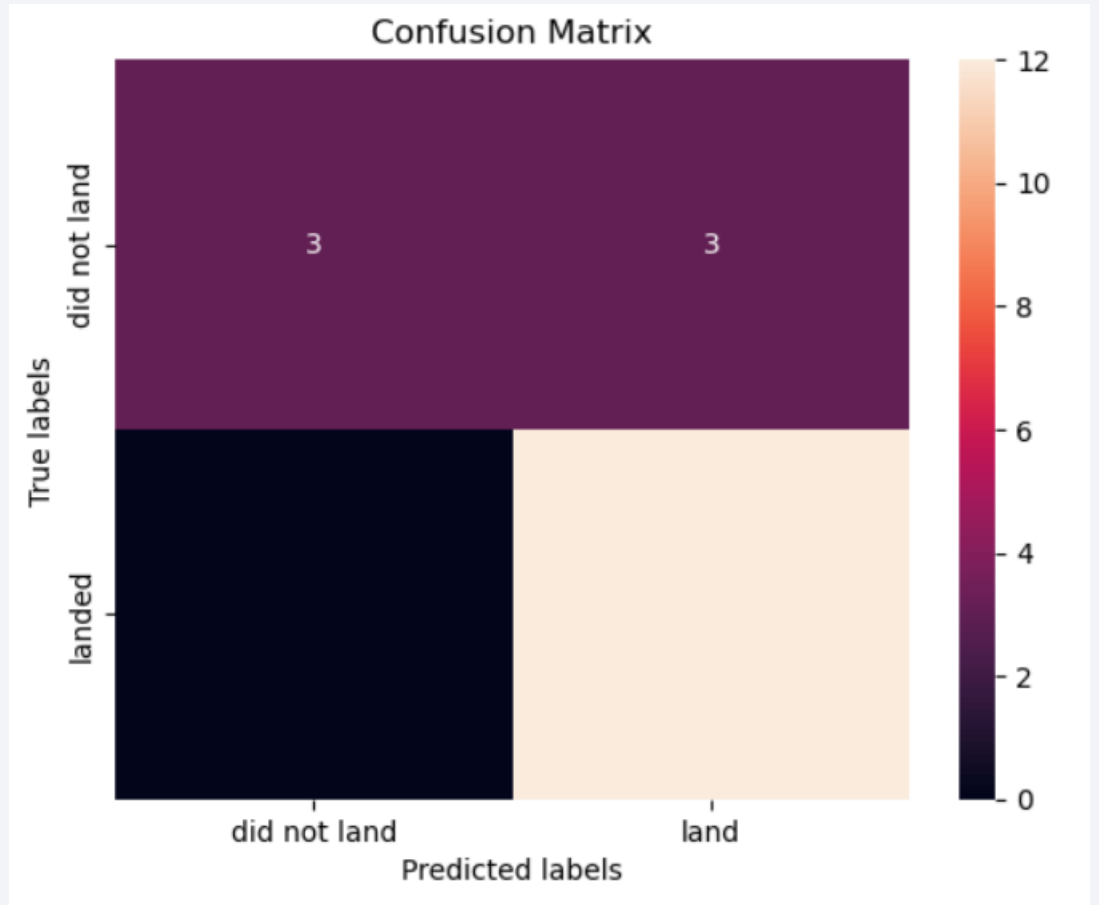
---

- After realizing a grid search for each model, the four models had a similar accuracy of 83.3% with the test dataset.



# Confusion Matrix

- The confusion matrix is similar for each model.
- True positive predictions are immaculate while there is a problem with false positive predictions.





# Conclusions

---

- Relevant features that influence a successful landing are Payload Mass, destination Orbit and Launch Site.
- Launch sites are located closer to the coasts and away from cities or fluent roads for safeness in case of failure.
- Launch success trend is growing positively through years.
- Booster landing prediction model reached an 83.3% accuracy, mostly guessing successful landings.

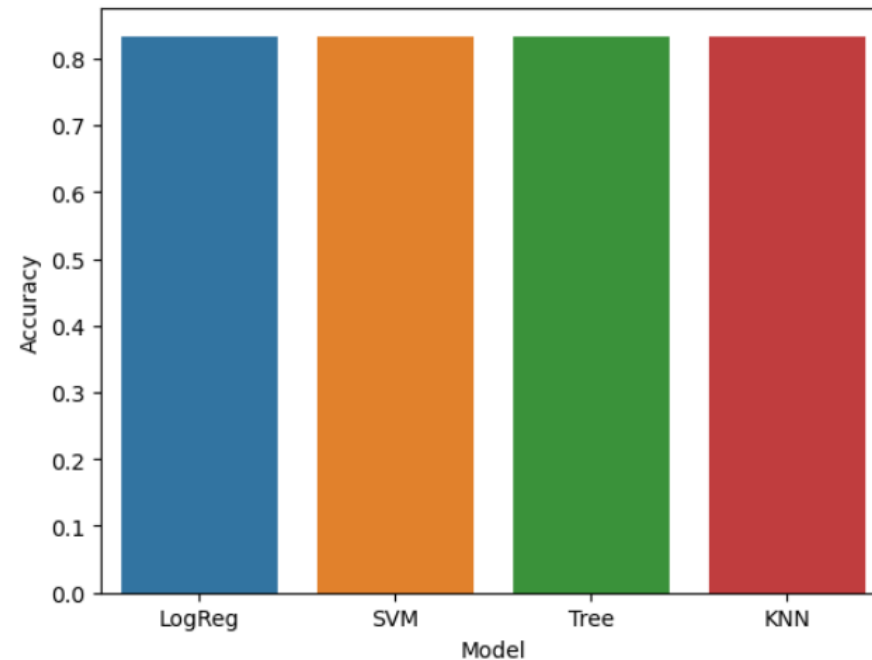
# Appendix

---

- This is the bar chart generated with seaborn for visualizing the accuracy data for each model

```
[33]: scores = pd.DataFrame(columns=['Model', 'Accuracy'],  
                           data={'Model': ['LogReg', 'SVM', 'Tree', 'KNN'],  
                                'Accuracy': [0.8333333333333334, 0.8333333333333334, 0.8333333333333334, 0.8333333333333334]})  
sns.barplot(scores, x='Model', y='Accuracy')
```

```
[33]: <Axes: xlabel='Model', ylabel='Accuracy'>
```



Thank you!

