



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

<Anshumaan Garg>  
<May, 2023>



# Outline

---

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

# Executive Summary

---

- Summary of methodologies

Various official SpaceX sources as well as Wikipedia's online data repository provides us with the data about the rocket launches and landing schemes. The data scraped in the form of html files(with the help of Beautiful Soup) as well as the data scraped and collected by SpaceX data collection centers has been used.

Exporting the data into an SQL database as well as a pandas data frame for analysis simplified the task of analyzing the data. It provides a precise methodology for analyzing the various launch sites and successful landing frequencies for the Falcon 9 stage 1 rockets.

# Executive Summary

---

- Summary of all results

Various prediction algorithms including SVM, KNN and multiclass prediction were used with the testing data as sample for predicting whether or not the model could predict whether the Falcon 9 stage 1 could be reused or not.

The SVM sigmoid classification has yielded an 83.33% success regarding the prediction of the target variable

# Introduction

---

- The project deals with the statistical problem for a space research company, using Falcon 9 stage 1 rocket launchers. The data has been taken from various sources such as Wikipedia Falcon 9 launch data as well as SpaceX data centers' data.
- The data provided, although labelled, still has a lot of dirt which needs to be cleaned in order to perform the secondary data analysis.
- The company wishes to predict whether the Falcon 9 stage 1 launchers can be reused based on the launch conditions and other independent variables that include the launch site, location, payload mass, etc.



Section 1

# Methodology

# Methodology

---

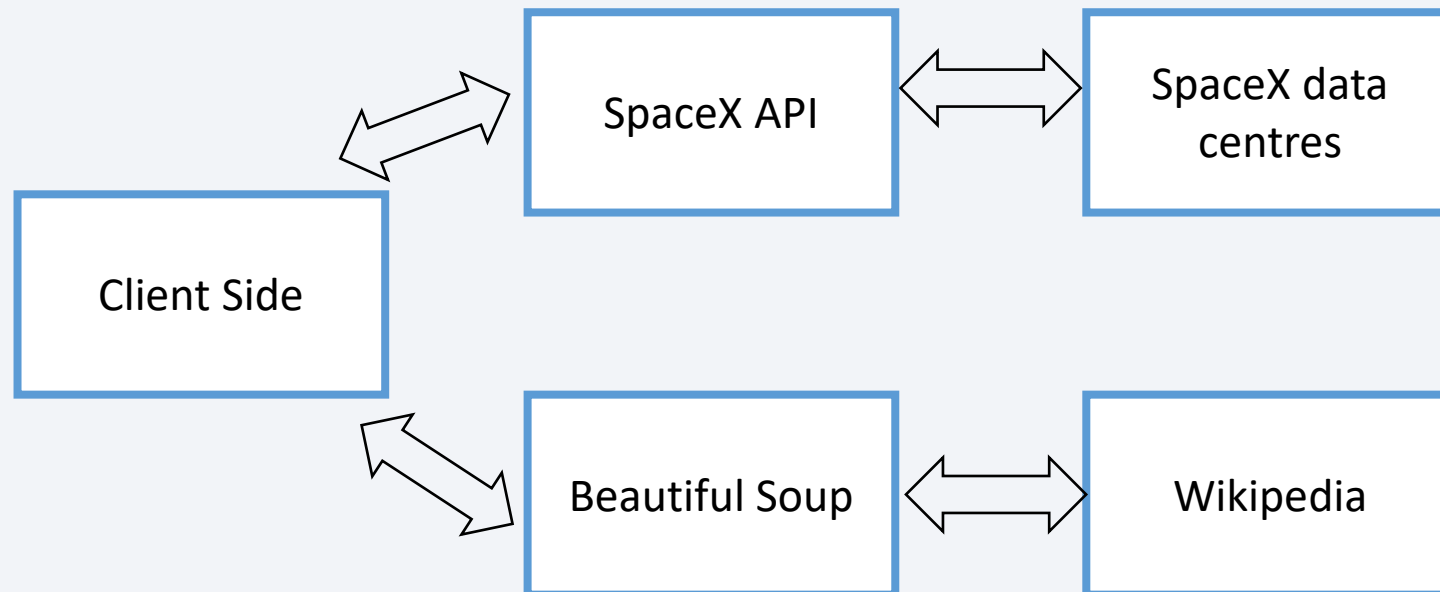
## Executive Summary

- Data collection methodology:
  - The data extraction of launch data with the help of SpaceX API and also Wikipedia Falcon 9 launch data (Beautiful Soup).
- Perform data wrangling
  - Analyzing the LABELLED but dirty data with the help of Pandas.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Various classification models such as SVM, Logistic Regression, etc. are used along with training and testing data sets.

# Data Collection

---

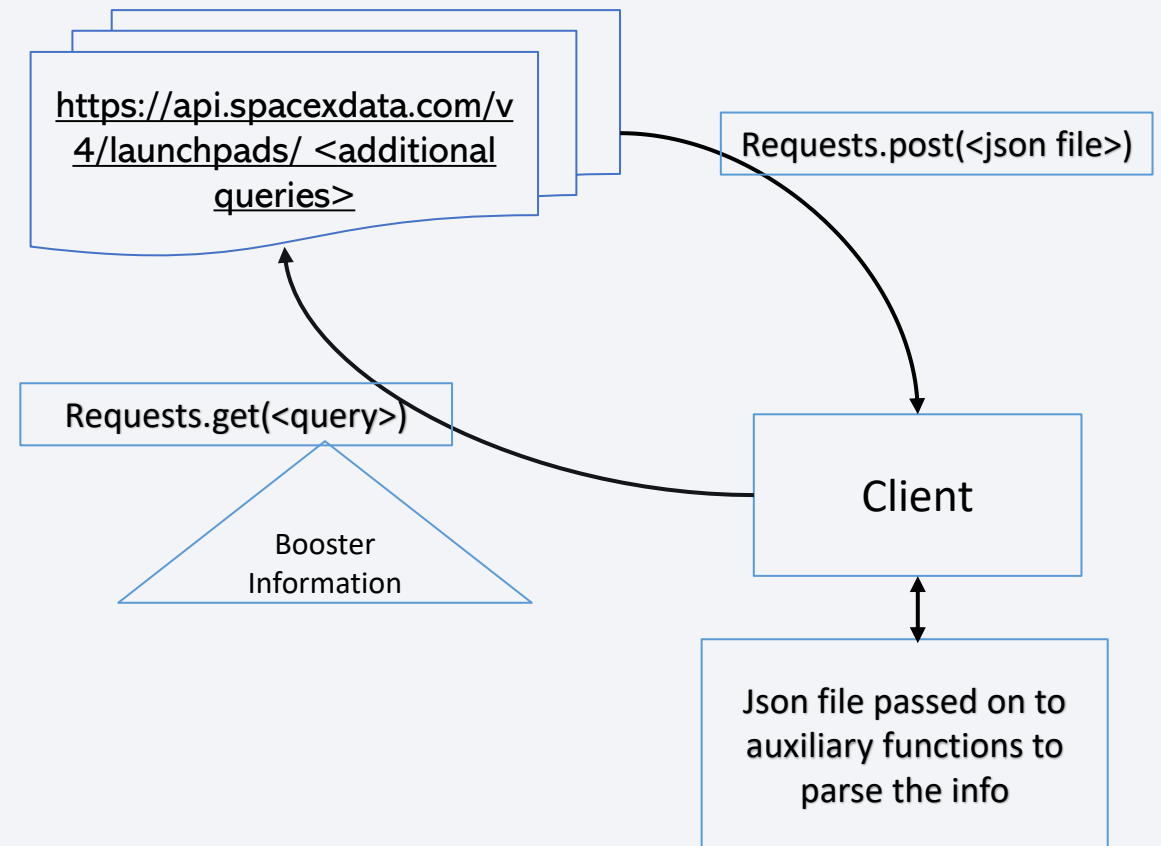
- Two main sources from where this data has been collected from are SpaceX API and





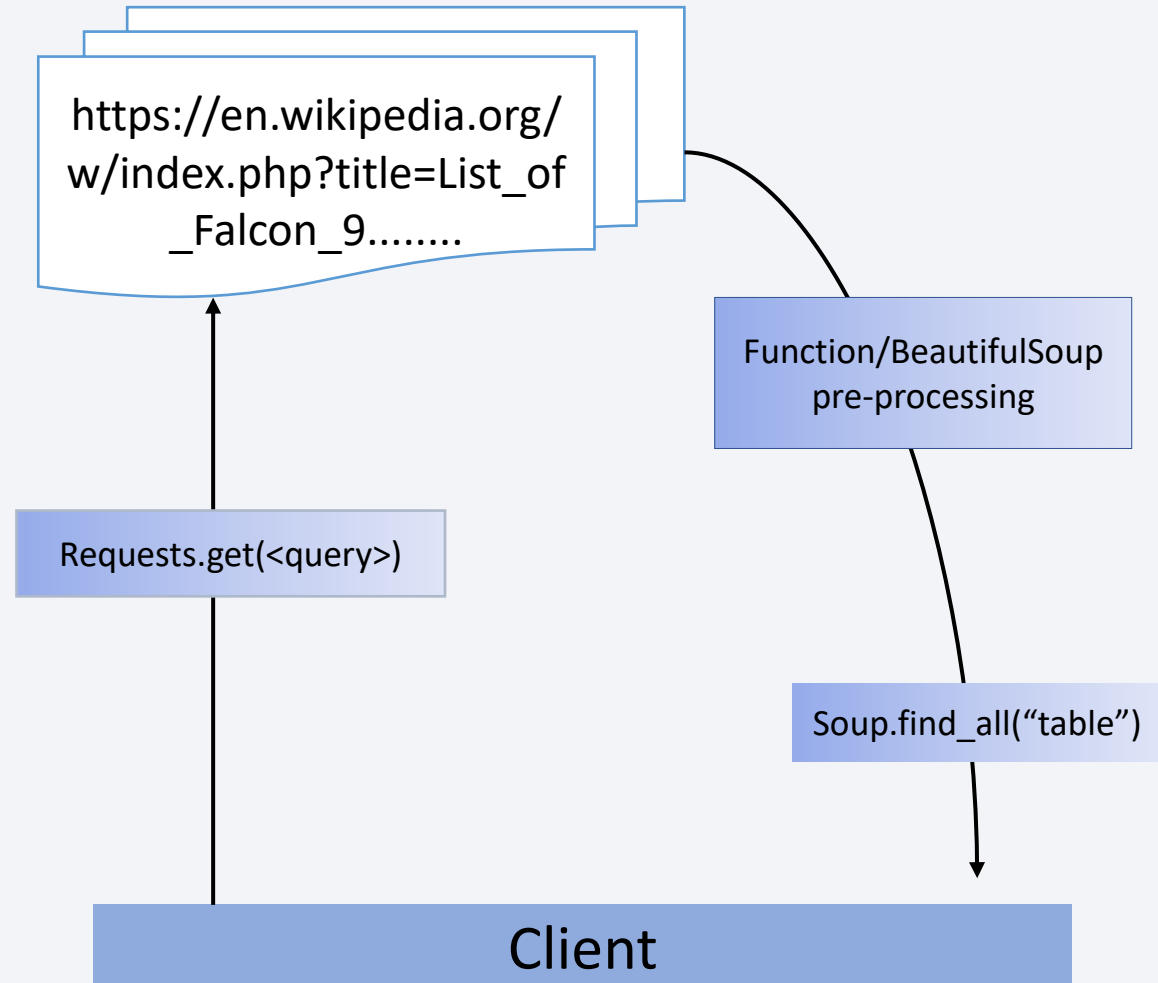
# Data Collection – SpaceX API

- SpaceX API works on the REST API for fetching and displaying data from various SpaceX data sites. The required data is fetched from the SpaceX API URL (<https://api.spacexdata.com/v4/launchpads/ <additional queries>>). Then a “GET” Rest request is posted to the server where the data is passed in the form of a .json file.
- [https://github.com/Anshumaan-Garg-20349/IBM\\_Final\\_Project.git](https://github.com/Anshumaan-Garg-20349/IBM_Final_Project.git)



# Data Collection - Scraping

- Falcon 9's booster launching and landing data can be scraped from the Wiki page with the help of a GET request and the html file can be scraped for data with the help of BeautifulSoup. The data is parsed as an html file where all the data is extracted from the table tags and stored in an external Pandas data frame.
- <[https://github.com/Anshumaan-Garg-20349/IBM\\_Final\\_Project.git](https://github.com/Anshumaan-Garg-20349/IBM_Final_Project.git)>



# Data Wrangling

---

- Since the data that was extracted from Wiki was labelled but not cleaned, the null, faulty-dtypes, or other erroneous values must be rectified. This can be done by the Pandas and Numpy library. The data frame is analyzed for the percentage of null values (`df.isnull().sum()/df.shape[0]*100`).
- The data frame is then checked for wrong datatypes such as a categorical attribute which needs to be an integer64. Thankfully this dataset has no dtype errors.
- <[https://github.com/Anshumaan-Garg-20349/IBM\\_Final\\_Project.git](https://github.com/Anshumaan-Garg-20349/IBM_Final_Project.git)>

## Data frame

### Pre-processing Magic / Data Wrangling

```
Dataframe.isnull().sum()/dataframe.shape[0]*100  
#for getting the percentage of null values in each  
column
```

```
Dataframe.dtypes  
#for analysing the data types of all the columns
```

# EDA with Data Visualization

---

<[https://github.com/Anshumaan-Garg-20349/IBM\\_Final\\_Project.git](https://github.com/Anshumaan-Garg-20349/IBM_Final_Project.git)>

- The data analysis steps to the EDA stage where SQL is used to query the meaningful data easily and Pandas and Matplotlib have been used for basic visualization of scatter plots between different independent variables to determine whether the success/failure depends on the corresponding parameters or not.

# EDA with SQL

- The database contains the data for 4 different launch sites for Falcon 9 boosters. The analysis performed lists:
  - `distinct(Launch_Sites)` for getting various launch statistics such as Payload capacity, orbit, and Landing\_Outcome, for each of these launch\_site.
  - Ascertaining the Success and Failure percentage for these launch sites and ranking them accordingly for which launch\_site has highest success ratio.
- <[https://github.com/Anshumaan-Garg-20349/IBM\\_Final\\_Project.git](https://github.com/Anshumaan-Garg-20349/IBM_Final_Project.git)>

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



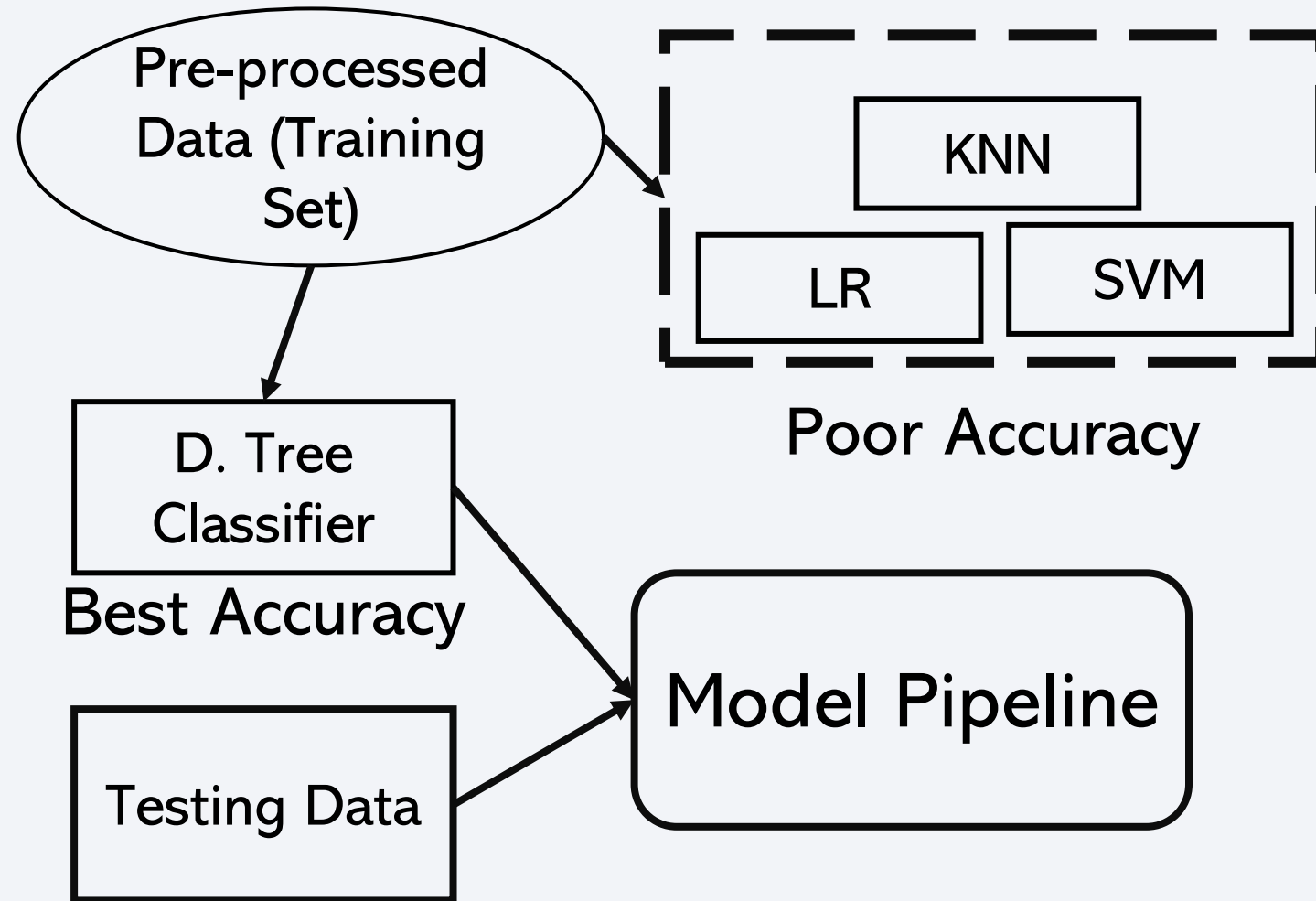
# Build an Interactive Map with Folium

---

- The Launch Sites have been displayed with pointers and little icons with green/red color indicating the success/failure have been clustered together for each of the individual launch site.
- The Pointers help locate the location of the launch site and uniquely identify them. The icon markers and circle markers denote the launch details.
- Link to project repository: <[https://github.com/Anshumaan-Garg-20349/IBM\\_Final\\_Project.git](https://github.com/Anshumaan-Garg-20349/IBM_Final_Project.git)>

# Predictive Analysis (Classification)

- The classifier model for the mission outcome at different launch sites has been evaluated with the help of 4 main classification algorithms.
- The best result was obtained with the Decision Tree Classification Algorithm with an accuracy score of 83.34%
- [https://github.com/Anshumaan-Garg-20349/IBM\\_Final\\_Project.git](https://github.com/Anshumaan-Garg-20349/IBM_Final_Project.git)





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.

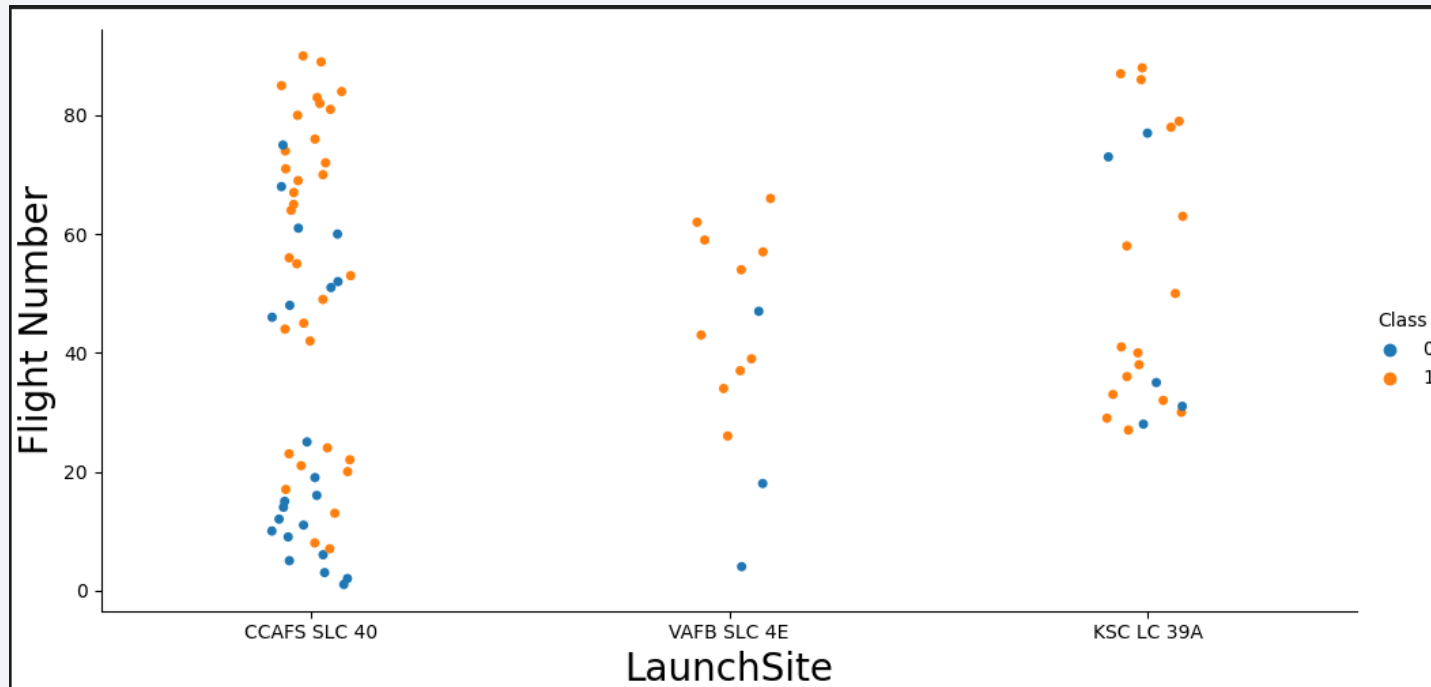
Section 2

# Insights drawn from EDA



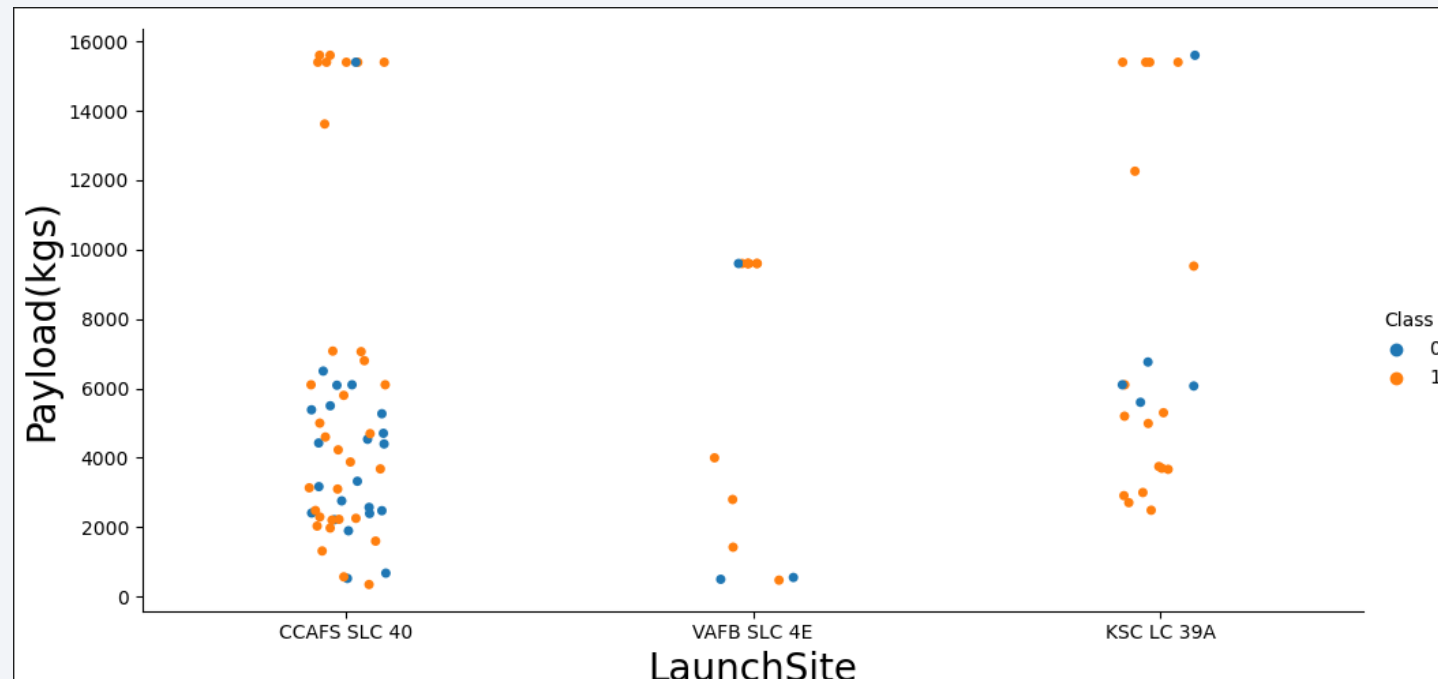
# Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site
- There is a strong relation between the Flight numbers and Launch Site location. As can be seen, the launch site “CCAFS SLC 40” has the highest successful launches (marked in orange), amongst all three.



# Payload vs. Launch Site

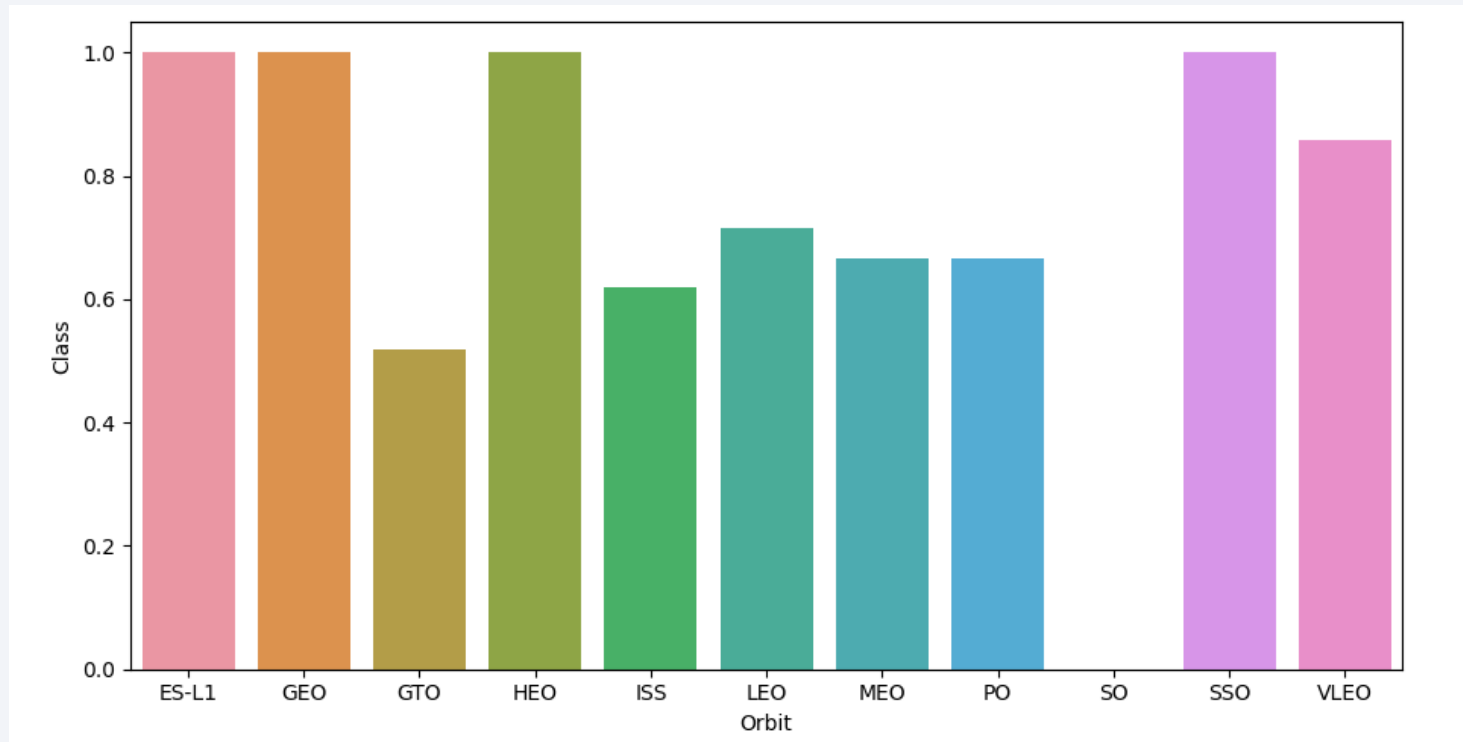
- Scatter plot of Payload vs. Launch Site
- Launch Site CCAFS SLC 40 has a high success ratio for heavy payloads whereas the site VAFB has no recorded launches for payload above 10k kgs.
- For heavy payloads(> 10,000kgs), the launch site CCAFS has the highest success ratio.





# Success Rate vs. Orbit Type

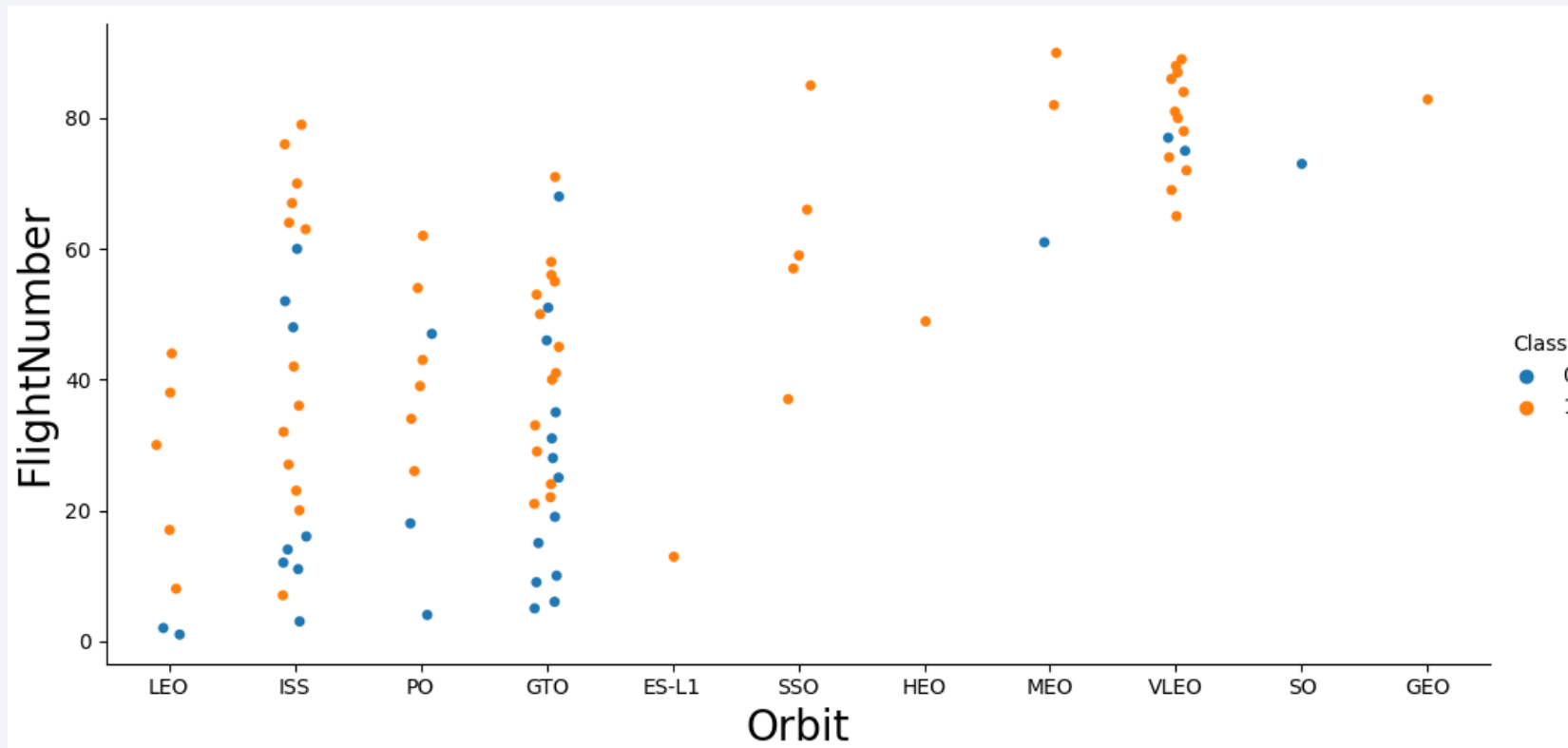
---



- Bar chart for the success rate of each orbit type.
- The orbit of the rocket plays a significant role in the success of the mission.
- ES-L1, GEO, HEO, SSO, VLEO have the highest success ratios.

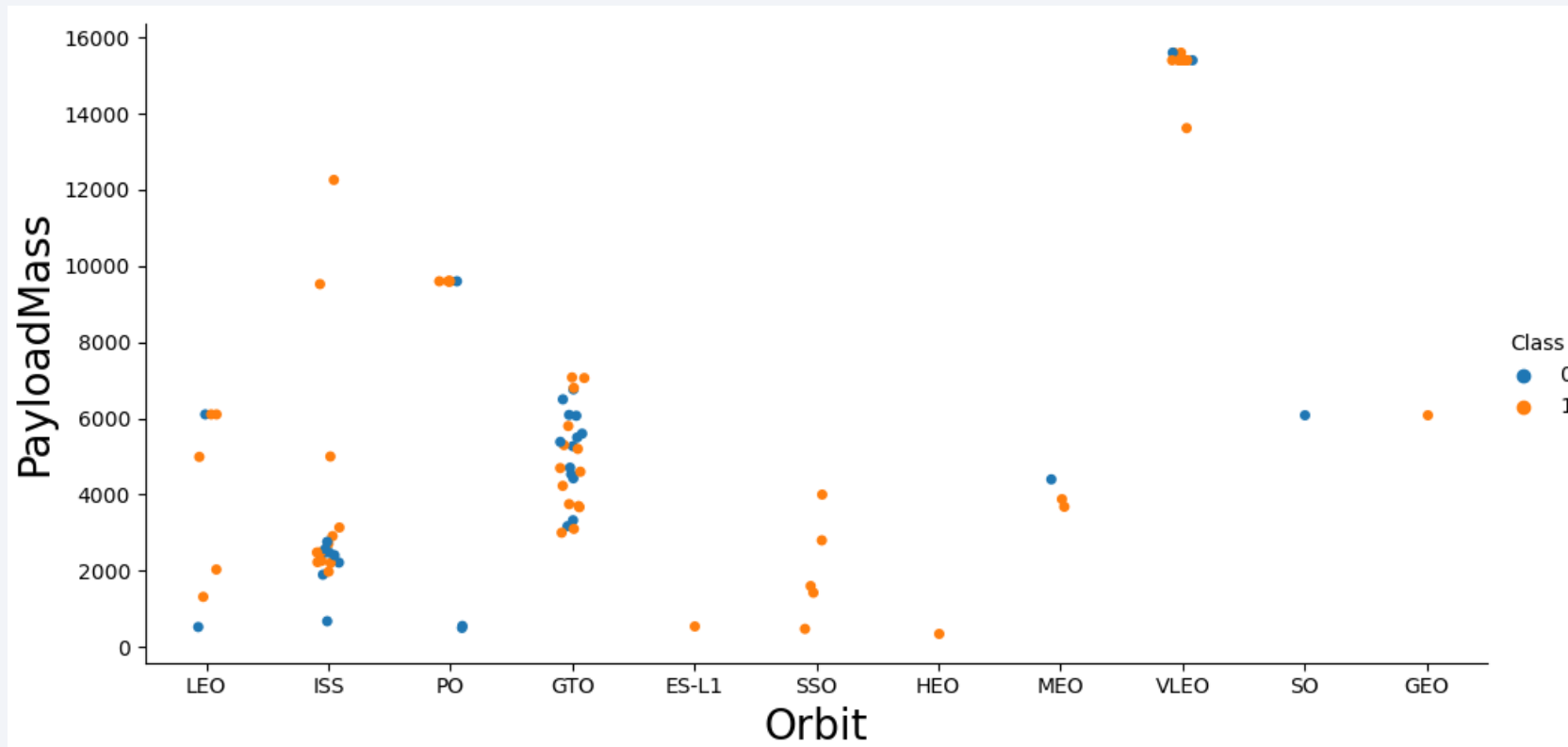
# Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type
- The Low Earth Orbit launches have the highest correlation with success ratio(dots in orange).



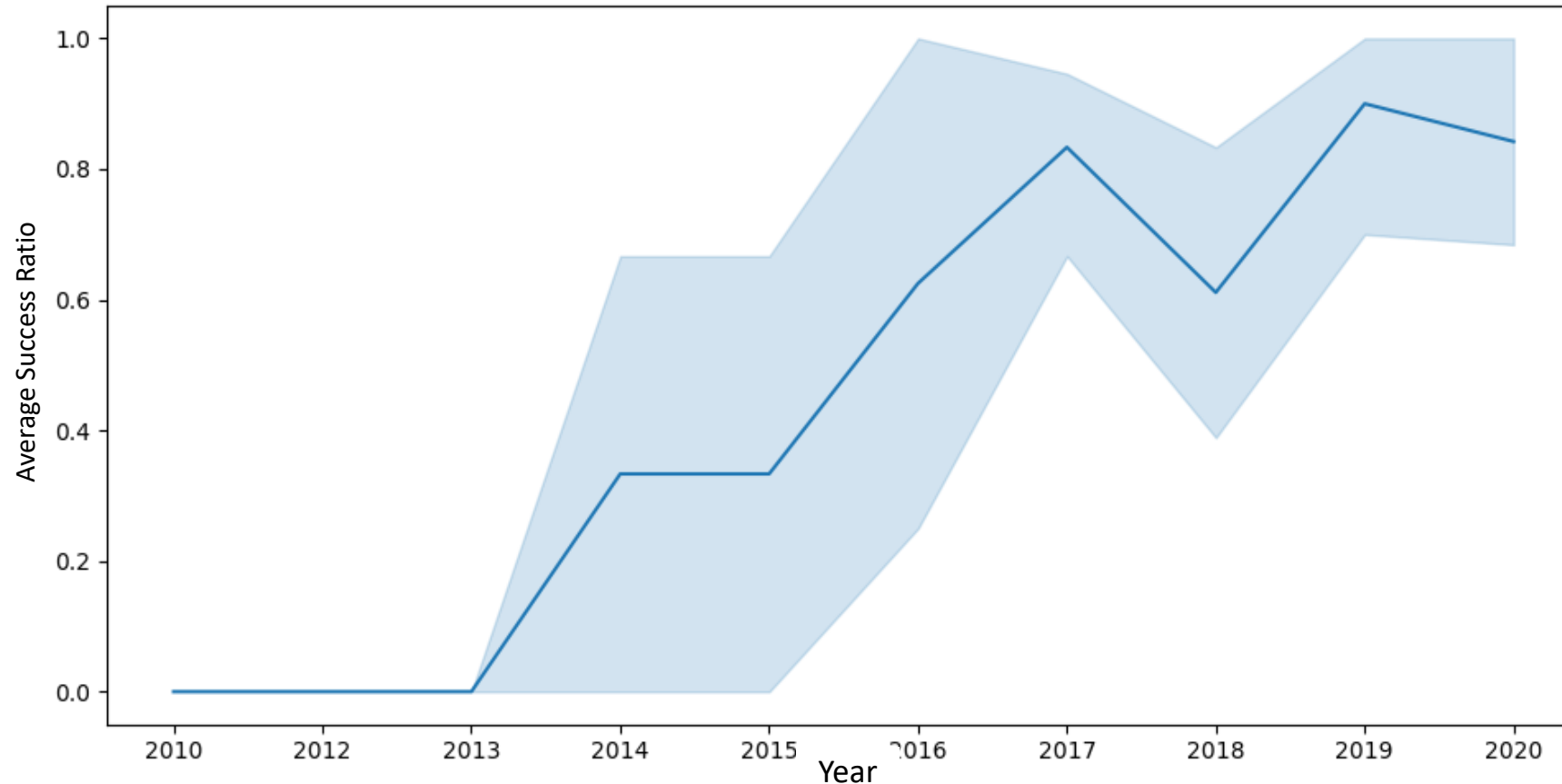
# Payload vs. Orbit Type

- Scatter point of payload vs. orbit type
- It can be interpreted from this that for payload(<8000kgs) the SSO, HEO have high success ratios, whereas for heavier payloads Polar orbit launches and ISS launches are more successful.



# Launch Success Yearly Trend

---



- Line chart of yearly average success rate
- The Average success rate for all Falcon 9 launches have seen a rise for the period 2013 to 2020.

# All Launch Site Names

---

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

The analysis of database with SQL lists all the distinct launch site names.



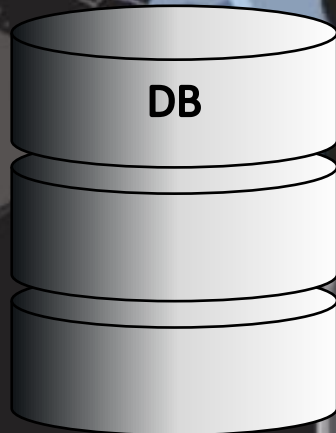


# Launch Site Names Begin with 'CCA'

- Some Significant Launches:

Launch Site	Payload	Customer
CCAFS LC-40	Dragon Spacecraft Qualification Unit	SpaceX
CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Bruere cheese	NASA (COTS) NRO
CCAFS LC-40	Dragon demo flight C2	NASA (COTS)
CCAFS LC-40	SpaceX CRS-1	NASA (CRS)
CCAFS LC-40	SpaceX CRS-2	NASA (CRS)

# Total Payload Mass



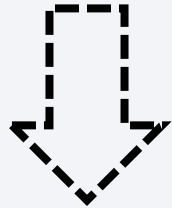
`select sum(PAYLOAD_MASS_KG_) as Total_Payload  
from spacextbl where customer like "NASA%";`

99980 kgs

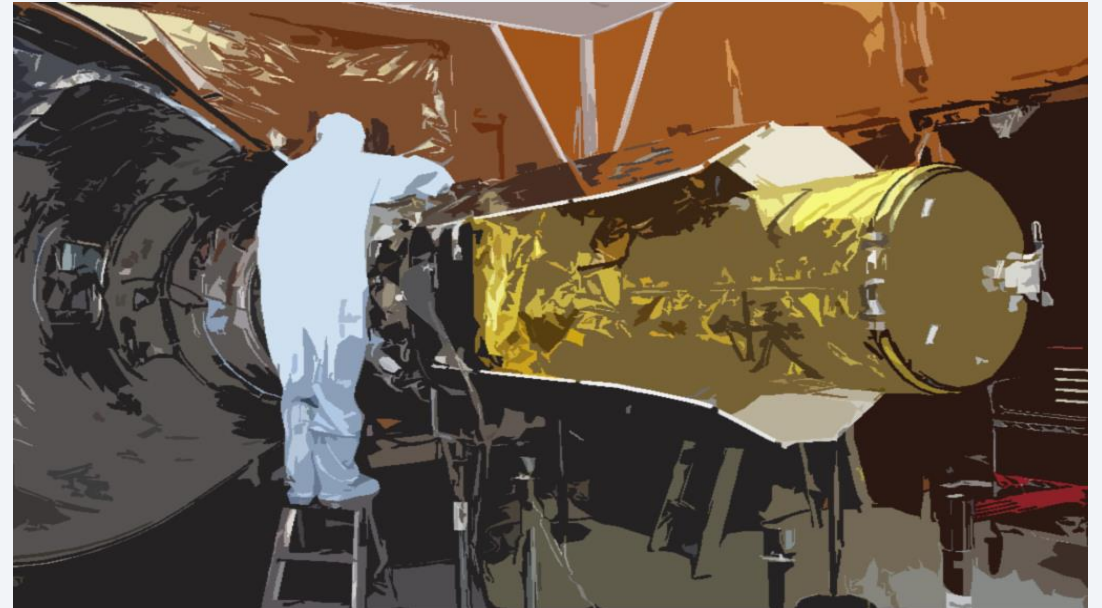
# Average Payload Mass by F9 v1.1

---


```
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL WHERE
BOOSTER_VERSION LIKE "F9 V1.1%";
```



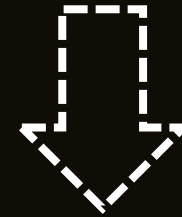
2534.667 kgs



# First Successful Ground Landing Date



SELECT MIN(DATE) FROM SPACEXTBL  
WHERE "LANDING\_OUTCOME" LIKE  
"%SUCCESS (GROUND PAD)%"



**THE BIG DAY!**  
**01-05-2017**



# Successful Drone Ship Landing with Payload between 4000 and 6000

---

## Booster\_Version

F9 FT B1021.1

F9 FT B1022

F9 FT B1023.1

F9 FT B1026

F9 FT B1029.1

F9 FT B1021.2

F9 FT B1029.2

F9 FT B1036.1

F9 FT B1038.1

F9 B4 B1041.1

F9 FT B1031.2

F9 B4 B1042.1

F9 B4 B1045.1

F9 B5 B1046.1

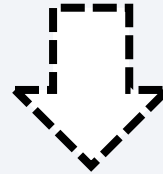
```
SELECT "BOOSTER_VERSION" FROM SPACEXTBL
WHERE "PAYLOAD_MASS_KG_" > 4000 &
"PAYLOAD_MASS_KG_" < 6000 AND "LANDING
_OUTCOME" LIKE "%SUCCESS (DRONE SHIP)%"
```



# Total Number of Successful and Failure Mission Outcomes

---

```
SELECT  
DISTINCT(MISSION_OUTCOME), COUNT(*)  
FROM SPACEXTBL GROUP BY  
MISSION_OUTCOME;
```



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

# Boosters Carried Maximum Payload

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

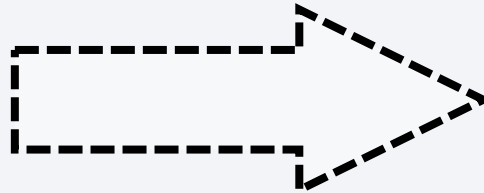
The maximum payload mass(in kgs) carried by Falcon 9 is 15600 kgs

```
SELECT BOOSTER_VERSION
FROM SPACEXTBL WHERE
PAYLOAD_MASS__KG_ =
(SELECT
MAX(PAYLOAD_MASS__KG_)
FROM SPACEXTBL)
```

# 2015 Launch Records

---

```
SELECT
DATE, BOOSTER_VERSION, LAUNCH
_SITE, "LANDING _OUTCOME"
FROM SPACEXTBL WHERE
SUBSTR(Date,7,4)='2015' AND
"LANDING _OUTCOME" LIKE
"%FAILURE%"
```



Date	Booster_Versi on	Launch_Sit e	Landing _Outcome
10-01- 2015	F9 v1.1 B1012	CCAFS LC- 40	Failure (drone ship)
14-04- 2015	F9 v1.1 B1015	CCAFS LC- 40	Failure (drone ship)

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

```
SELECT DISTINCT("LANDING
_OUTCOME"), COUNT(*) FROM
SPACEXTBL WHERE "LANDING _OUTCOME"
LIKE "%SUCCESS%" AND DATE BETWEEN
'04-06-2010' AND '20-03-2017'
GROUP BY "LANDING _OUTCOME" ORDER
BY "COUNT(*) DESC"
```

Landing _Outcome	No. of Launches
Success	20
Success (drone ship)	8
Success (ground pad)	6

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.

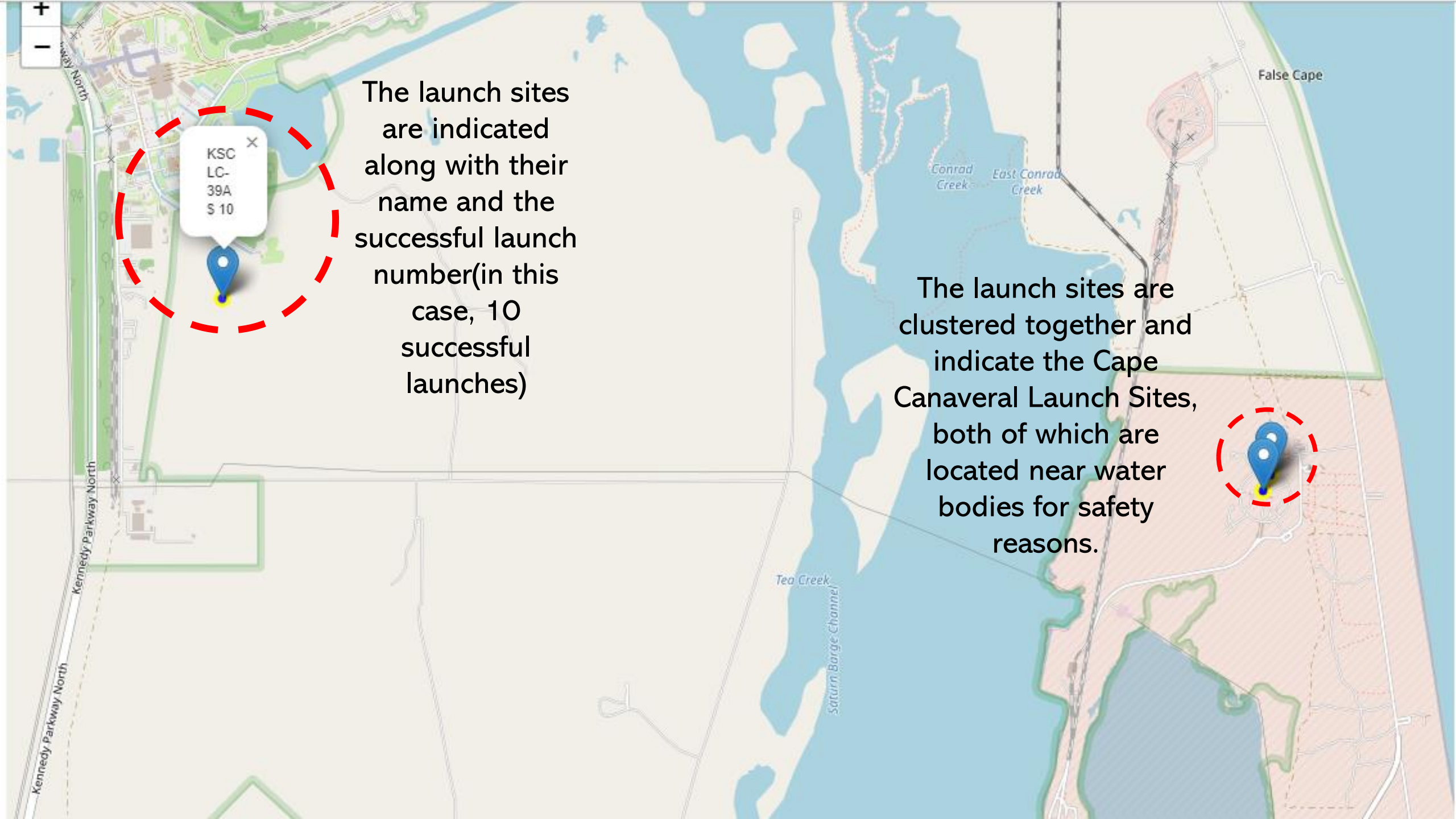
Section 3

# Launch Sites Proximities Analysis



# Launch Sites' Loc(World Map)

- The map shows the location of the Launch Sites under study, marked in blue and yellow circles.
- The following illustration clearly shows that all the launch sites are in close proximity of a water body.



The launch sites are indicated along with their name and the successful launch number(in this case, 10 successful launches)

The launch sites are clustered together and indicate the Cape Canaveral Launch Sites, both of which are located near water bodies for safety reasons.



# Launch Clusters for Launch Sites

- The following illustration shows the number of launches of each individual launch site clustered together.





The launches are marked around the launch site as a spiral with their colour indicating the outcome of the launch.

For this particular launch site, it can be seen that most of the launches were successful except three.

The number indicates the number of launches that are clustered at that particular launch site.





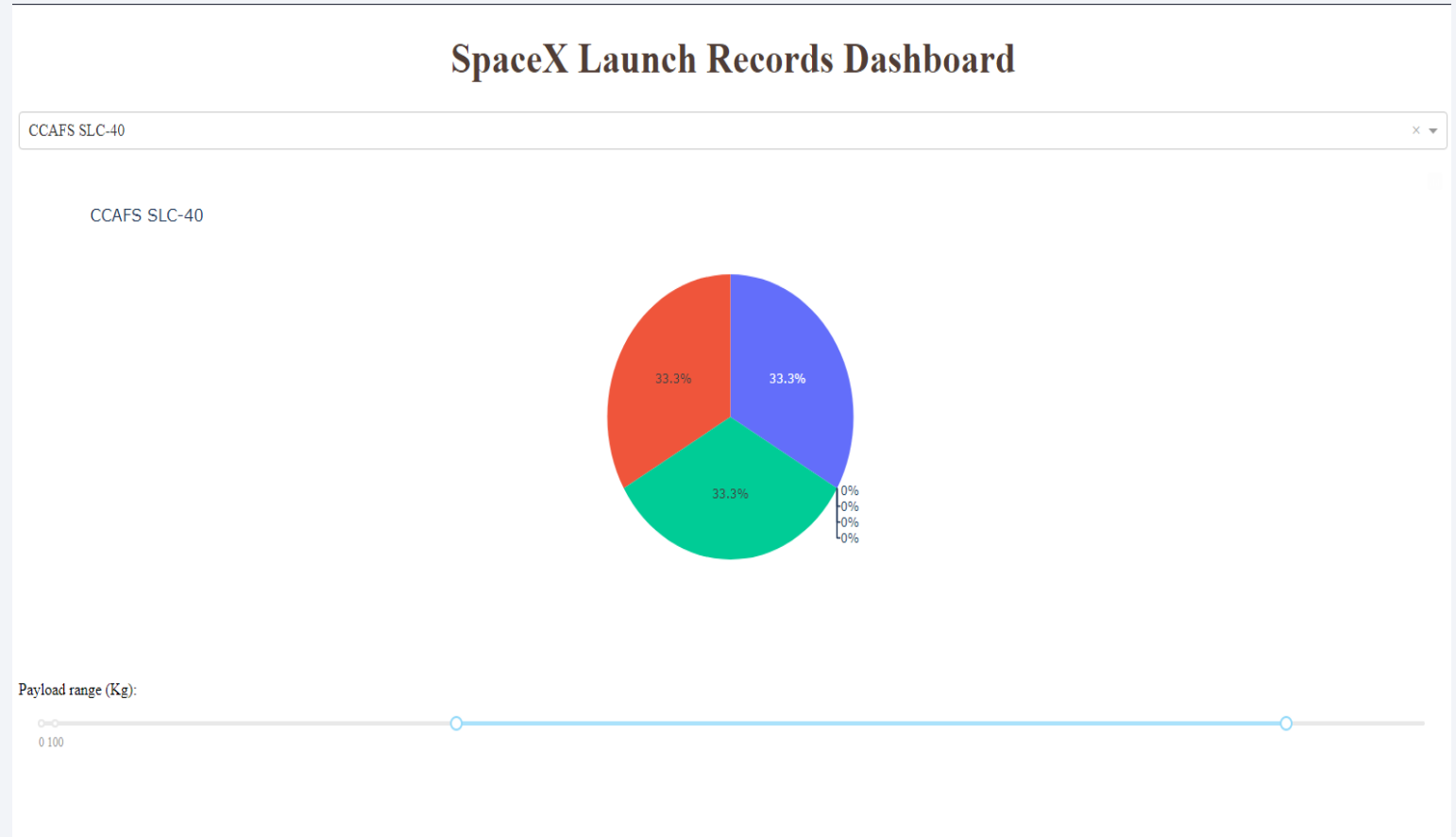
Section 4

# Build a Dashboard with Plotly Dash

# Dashboard pie chart representation

The success/failure ratio of different launch sites can be viewed with the help of a web application.

The name of the sites can be selected from the drop down menu and the payload can be selected with the help of the range slider.

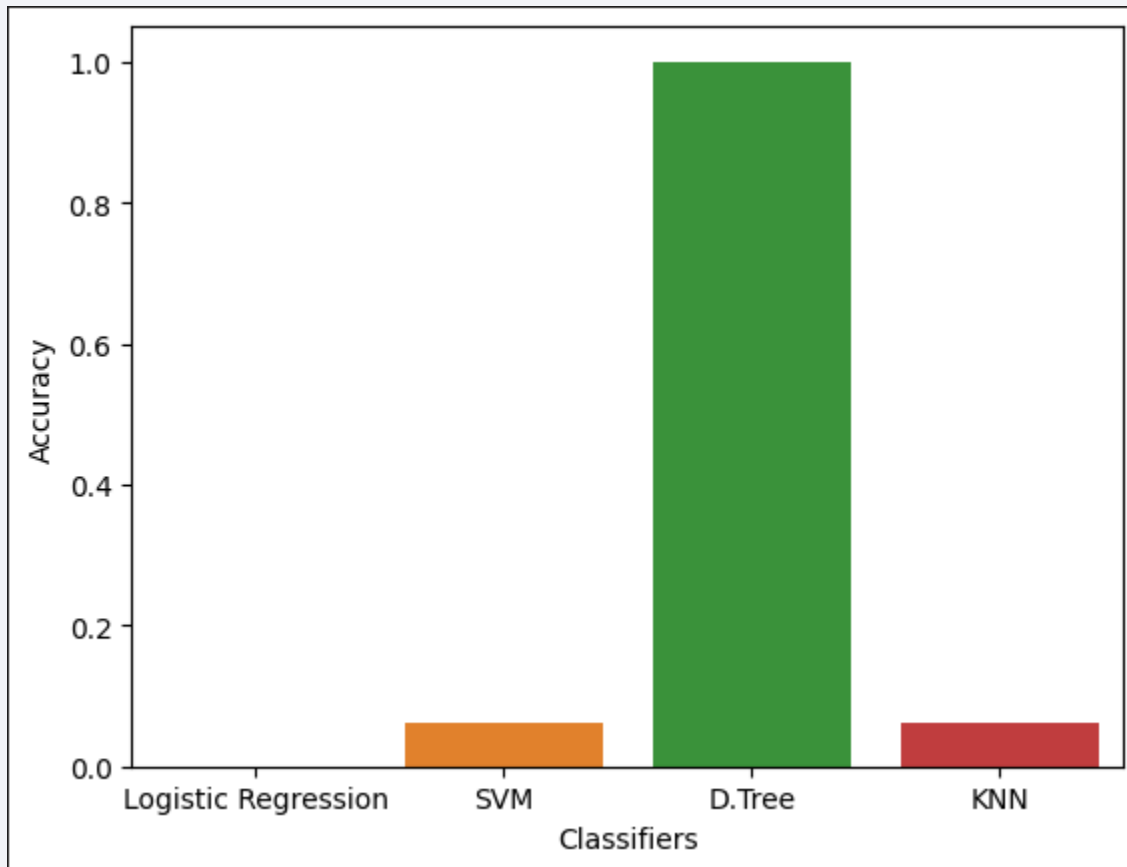


Section 5

# Predictive Analysis (Classification)

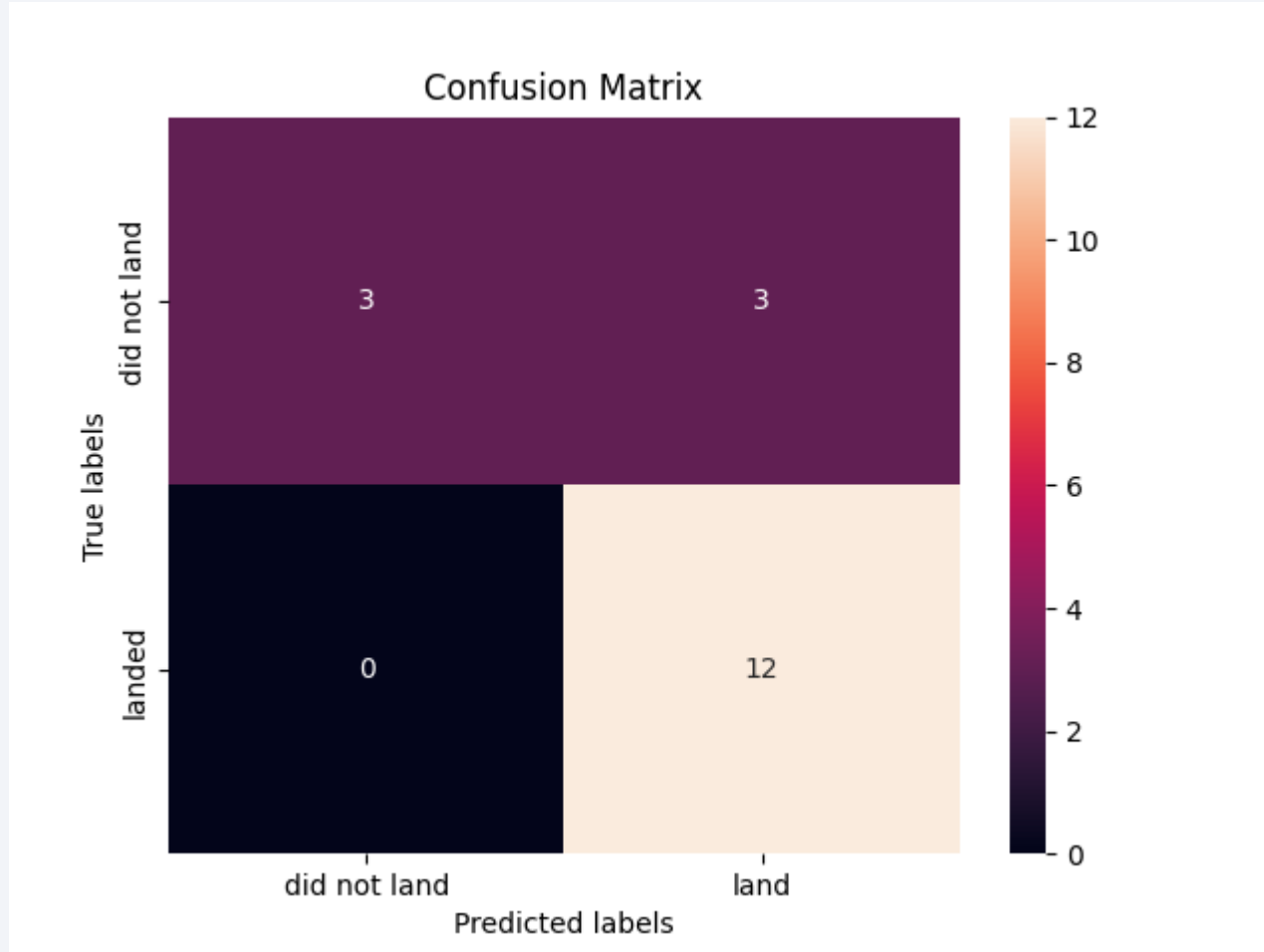
# Classification Accuracy

---



- The following bar plot shows the accuracy of different classification techniques normalized using Min-Max normalization technique.
- As is evident from the plot, Decision Tree Classification has the highest accuracy score.

# Confusion Matrix



- This is the confusion matrix of the Decision Tree Classifier.
- The model stands with the highest accuracy with an accuracy of 83.34%, with the false classification being for the 3 false positives.



# Conclusions

---

- The data that is provided to us can be used to build a Decision Tree Classifier for predicting the reusability of the Falcon 9 stage 1 boosters.
- The model uses various parameters as classifiers such as 'FlightNumber', 'PayloadMass', 'Flights', 'Block', 'ReusedCount', and other categorical dummies.
- Various models were used to evaluate the classification results and the classifier was tested on a training and testing set, where the Decision Tree Classifier shows the best results with an accuracy of 83.34% while predicting the reusability of Stage 1 Boosters.

# Appendix

---

- Various function for normalization of data in predictive analysis (refer to GIT Link)
- Seaborn plots: Accuracy-analysis of various models.

Thank you!

