

# Winning Space Race with Data Science

Thomas Mathew Panicker  
28<sup>th</sup> July 2024



# Outline

---

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

# Executive Summary

---

- Summary of methodologies
  - Data was sourced from the SpaceX public API and Wikipedia. Launch outcome information was extracted to serve as the dependent variable for Machine Learning models.
  - SQL queries and various visualizations (static plots, interactive maps, and a dashboard) were created to analyze the dataset and answer key questions.
  - Predictive models used included Logistic Regression, SVM, Decision Tree, and KNN.
- Summary of all results
  - Launch data includes details such as flight number, launch date, payload mass, orbit type, launch site, mission outcome, and other variables.
  - KNN showed highest accuracy and Logistic Regression, SVM showed comparable performance for Machine Learning models on this dataset with decision tree the worst ML model performance.

# Introduction

---

- SpaceX is one of the successful American aerospace manufacturer and space transportation company which tries to reduce the cost of space travel by using reusable first stage. Therefore if it can be determined whether the first stage will land, then it is possible to determine the cost of the launch.
- Space Y is a competitor of SpaceX and this project intends to find out the price of each launch by gathering data, preparing dashboards and determining if SpaceX will reuse stage one.
- This will be accomplished by training a machine learning model and using public information to predict if SpaceX will resuse the first stage.

Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection was done for the project by using SpaceX launch data collected through SpaceX REST API.
- Data wrangling - Data was cleaned to prepare for visualizations, queries, and Machine Learning models.
- Exploratory data analysis (EDA) was conducted using visualizations and SQL
- Interactive visual analytics were developed with Folium and Plotly Dash.
- Predictive analysis was performed using classification models.



# Data Collection

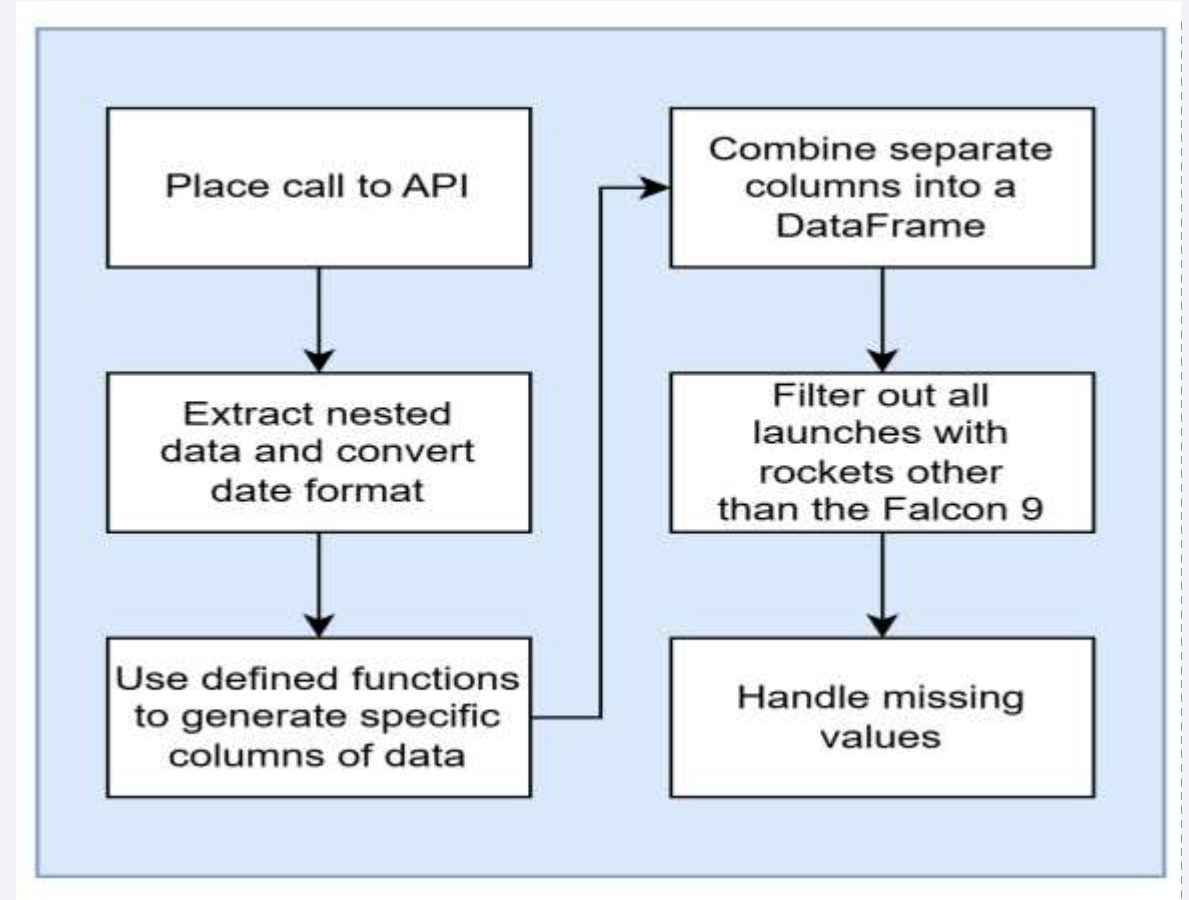
---

- Data is gathered for the project by using SpaceX launch data collected through SpaceX REST API. This API gives data about launches, type of rocket used, payload delivered, launch specifications, landing specifications and landing outcome. This data is the used to predict if SpaceX will attempt to land the rocket or not.
- <https://api.spacexdata.com/v4/>
- A get request is initiated to obtain the data from the various endpoints of the above API and receiving responses in JSON objects format. The `json_normalize` function is used to “normalize” the structured json data into a flat table.

# Data Collection – SpaceX API

- Data collected from publicly available data of SpaceX using GET request through the Space X API is converted into Pandas dataframe.
- GitHub URL of the completed SpaceX API calls notebook:

[testrepo/jupyter-labs-spacex-data-collection-api.ipynb](https://github.com/Mathewtmp/testrepo/blob/main/jupyter-labs-spacex-data-collection-api.ipynb) at main · Mathewtmp/testrepo (github.com)






# Data Collection - Scraping

- Another data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. These tables are scraped using beautiful soup and placed into pandas dataframe.

- Add the GitHub URL of the completed web scraping notebook:

[testrepo/jupyter-labs-webscraping.ipynb](https://testrepo/jupyter-labs-webscraping.ipynb) at [main · Mathewtmp/testrepo \(github.com\)](https://github.com/Mathewtmp/testrepo)

## Web scraping Falcon 9 Launch records



Web scraping with BeautifulSoup

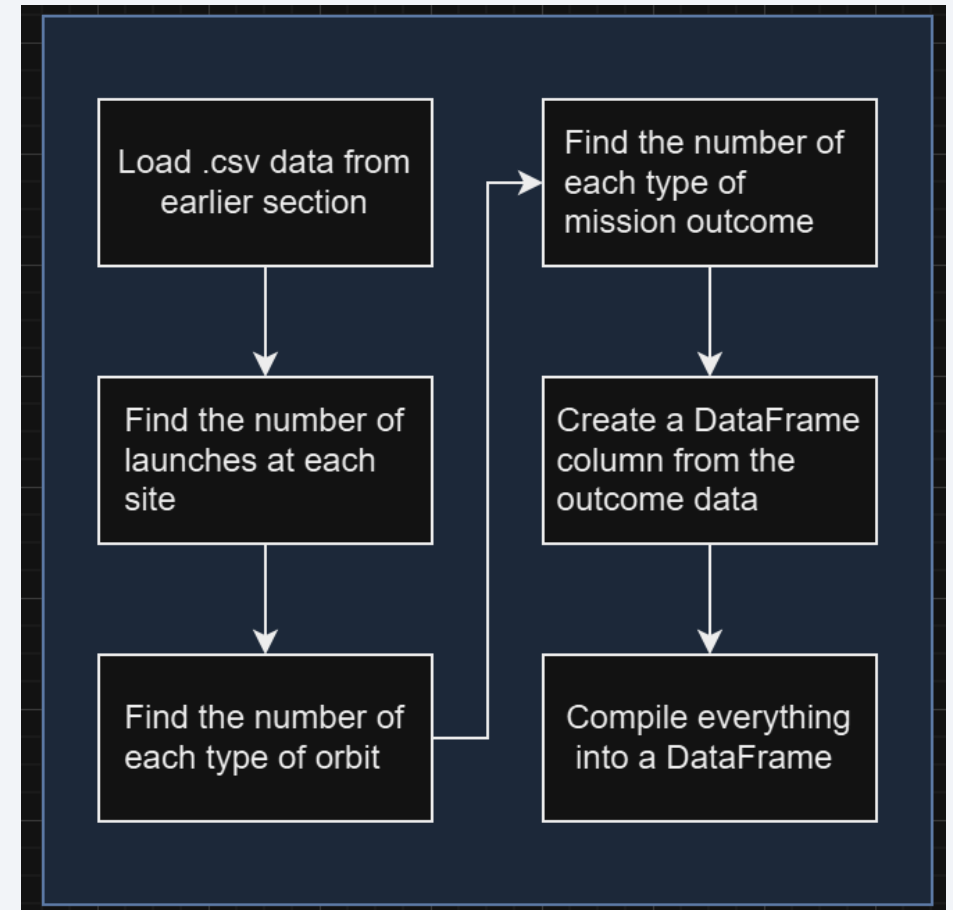
FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude		
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None	None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None	None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None	None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None	None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-08-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None	None	1	False	False	False	None	1.0	0	B0003	-80.577366	16.591857

# Data Wrangling

---

- The initial .csv file required data cleaning.
- Launch sites, orbit types, and mission outcomes were standardized.
- Mission outcomes were simplified to a binary format: 1 for a successful Falcon 9 first stage landing and 0 for a failure.
- This new binary classification was added to the DataFrame for further analysis.
- GitHub URL of the completed data wrangling related notebooks:

[testrepo/labs-jupyter-spacex-Data-wrangling.ipynb at main · Mathewtmp/testrepo \(github.com\)](#)



# EDA with Data Visualization

---

- Following charts were generated to analyze Launch Site trends:
  - Scatterplot showing the relationship between mission outcomes, Launch Sites, and Flight Numbers.
  - Scatterplot depicting the relationship between mission outcomes, Launch Sites, and Payloads.
- Following charts were created to examine Orbit Type trends:
  - Bar chart illustrating the relationship between mission outcomes and Orbit Types.
  - Scatterplot showing the relationship between mission outcomes, Orbit Types, and Flight Numbers.
  - Scatterplot depicting the relationship between mission outcomes, Orbit Types, and Payloads.
- A chart was created to observe trends over time:
  - Line plot showing the mission outcome trends by year.
- GitHub URL of the completed EDA with data visualization notebook:

[testrepo/edadataviz.ipynb at main · Mathewtmp/testrepo \(github.com\)](https://github.com/Mathewtmp/testrepo/blob/main/edadataviz.ipynb)

# EDA with SQL

---

- The following SQL queries you performed:
  - Display the names of unique launch sites used in the space missions.
  - Show 5 records where launch sites start with 'KSC'.
  - Show the total payload mass carried by boosters launched by NASA (CRS).
  - Show the average payload mass carried by the booster version F9 v1.1.
  - List the dates when successful landings were achieved on the drone ship.
  - List the names of successful boosters on the ground pad with payload masses between 4000 and 6000.
  - List the total number of successful and failed mission outcomes.
  - List the booster versions that carried the maximum payload mass.
  - Show records with month names, successful ground pad landings, booster versions, and launch sites for 2017.
  - Rank the count of successful landings between June 4, 2010, and March 20, 2017, in descending order.

# Build an Interactive Map with Folium

---

- Markers were added for launch sites and for the NASA Johnson Space Center
- Circles were added for the launch sites.
- Lines were added to show the distance to the nearby features:
- Distance from CCAFS LC-40 to the coastline
- Distance from CCAFS LC-40 to the rail line
- Distance from CCAFS LC-40 to the perimeter road
- GitHub URL of the completed interactive map with Folium map:

[testrepo/lab\\_jupyter\\_launch\\_site\\_location.ipynb](https://github.com/Mathewtmp/testrepo/blob/main/lab_jupyter_launch_site_location.ipynb) at main · Mathewtmp/testrepo (github.com)

# Build a Dashboard with Plotly Dash

---

- The input dropdown serves as a tool for choosing either a single or all launch sites for the pie chart and scatterplot. The pie chart then illustrates one of two scenarios:
- When 'All Sites' is selected, it shows the spread of successful Falcon 9 first stage landings across the sites.
- When a specific site is chosen, it presents the ratio of successful to unsuccessful Falcon 9 first stage landings at that particular site.
- The input slider functions as a filter for the payload masses that are displayed on the scatterplot. The scatterplot itself portrays the distribution of Falcon 9 first stage landings, divided by payload mass, mission result, and booster version category.
- GitHub URL of the completed Plotly Dash lab:

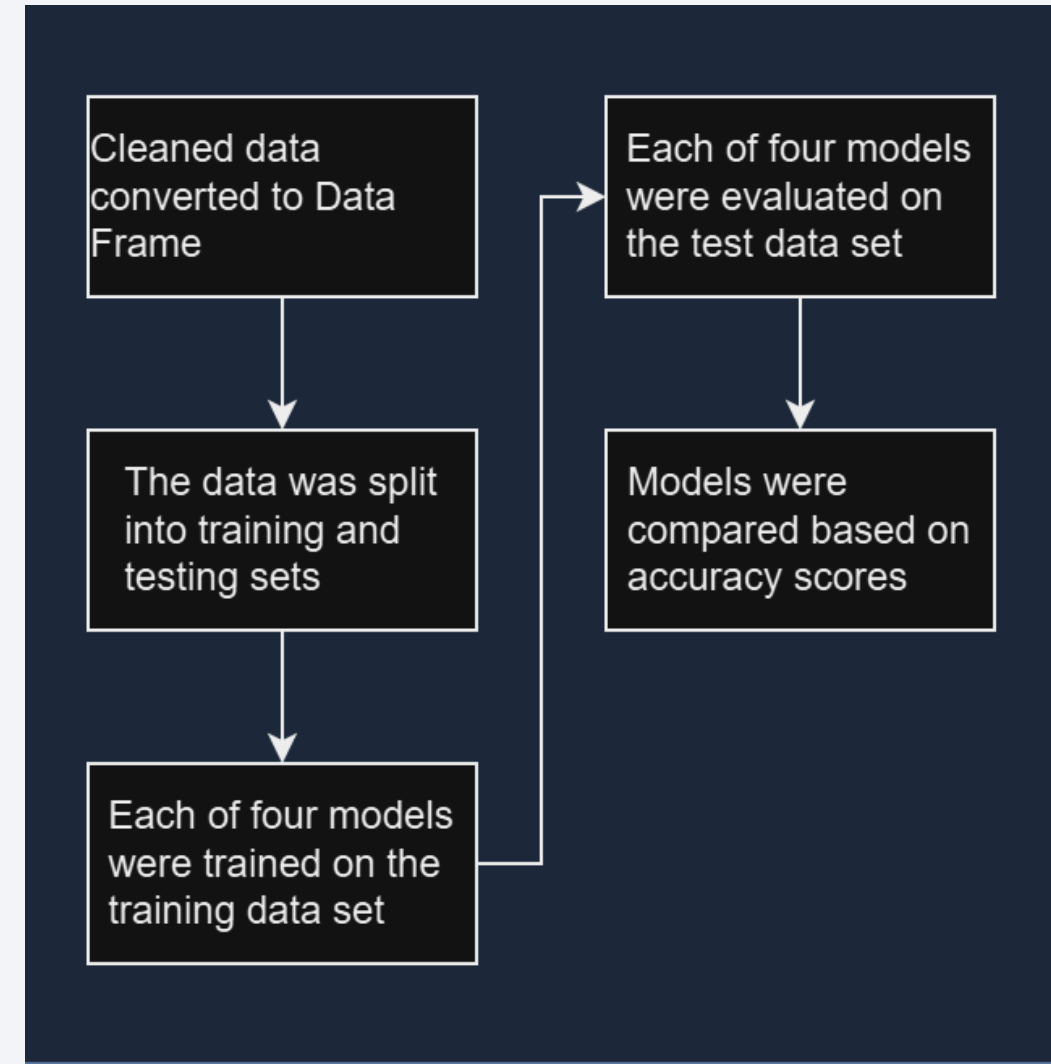
[testrepo/Plotly\\_Dash\\_Lab at main · Mathewtmp/testrepo \(github.com\)](https://github.com/testrepo/Plotly_Dash_Lab)



# Predictive Analysis (Classification)

- The dataset was divided into two parts: training and testing sets.
- Four machine learning models - Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN) - were trained using the training dataset.
- The evaluation of hyper-parameters was carried out using the GridSearchCV() function, and the optimal ones were chosen using the '.best\_params\_' attribute.
- With these optimal hyper-parameters, each of the four models' accuracy was scored using the testing dataset.
- GitHub URL of the completed predictive analysis lab:

[testrepo/SpaceX Machine Learning Prediction Part 5.ipynb at main · Mathewtmp/testrepo \(github.com\)](https://github.com/Mathewtmp/testrepo/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.ipynb)



# Results

---

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

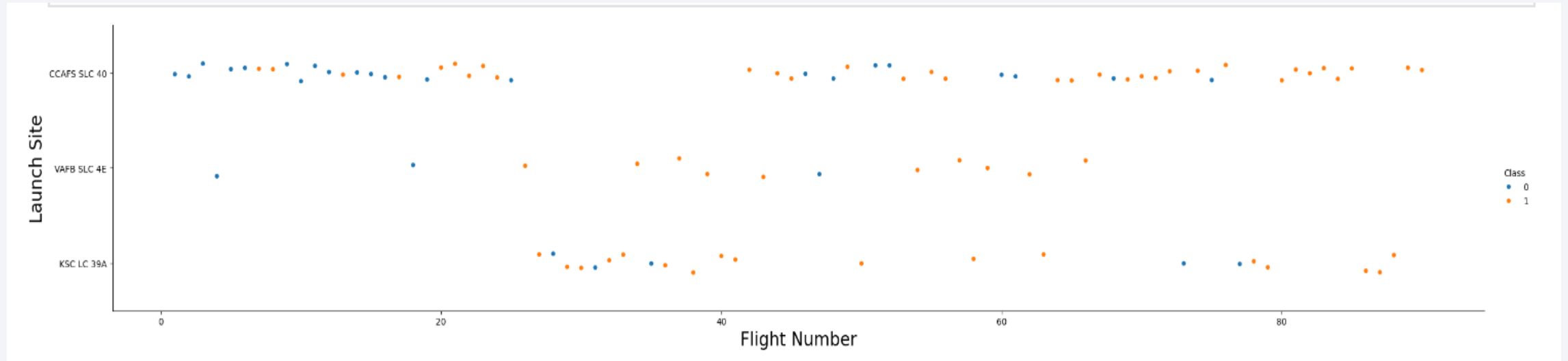
The background of the slide is an abstract composition. It features a solid blue area on the left side, which transitions into a dynamic pattern of diagonal streaks in shades of blue, red, and cyan on the right. A faint, light-blue grid pattern is visible across the entire image, particularly prominent in the blue and cyan areas.

Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site

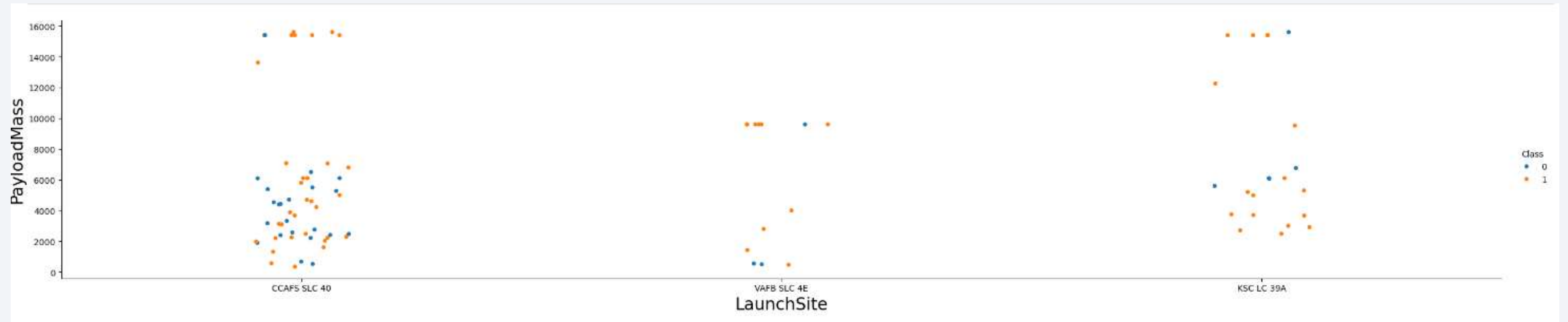


- Successful landings marked by Class 1 appears to increase as the flight numbers increase and CCAFS SLC 40 has most number of successful launches



# Payload vs. Launch Site

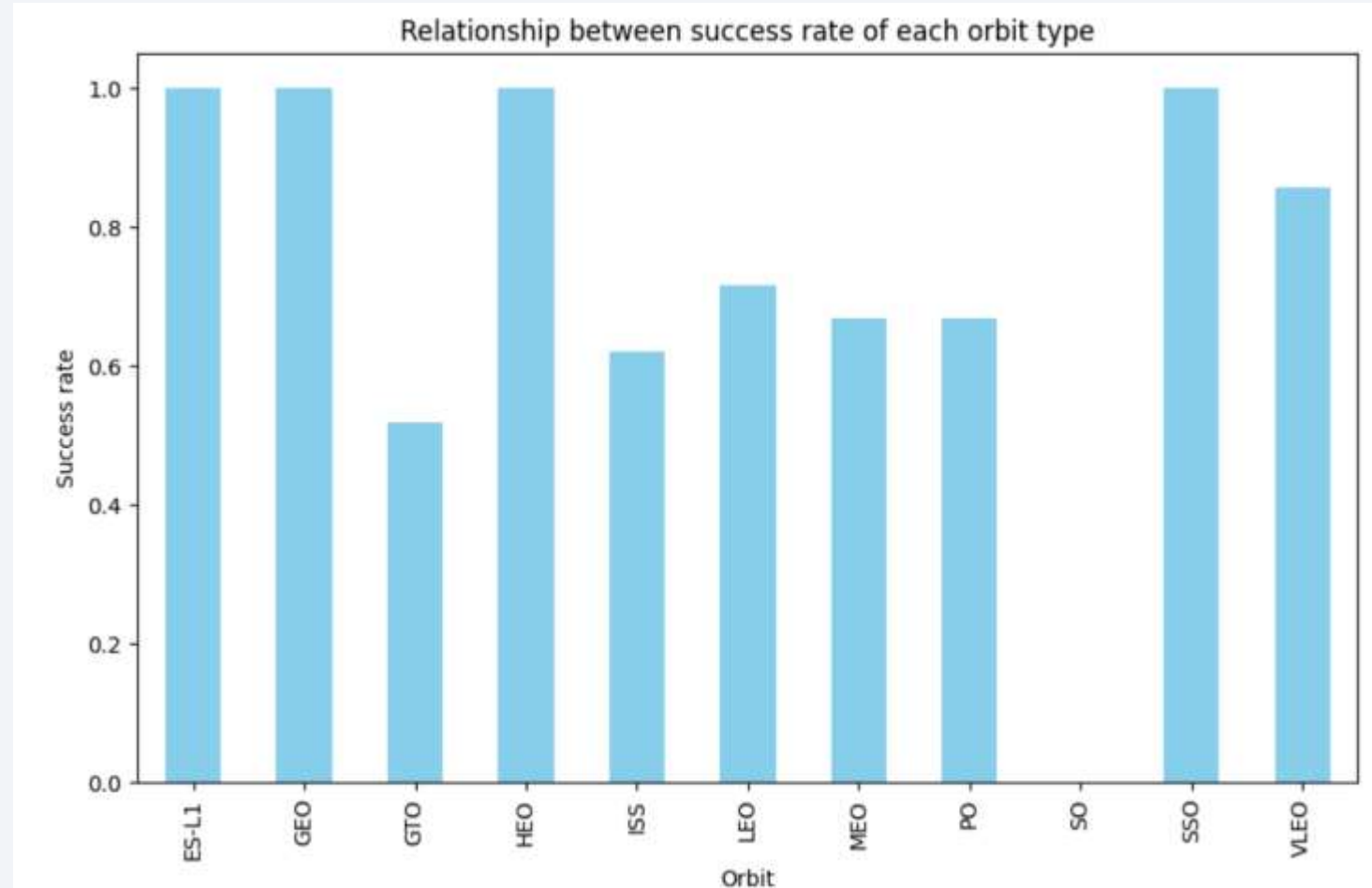
- Scatter plot of Payload vs. Launch Site



- CCAFS SLC 40 has most number of successful launches with lower PayloadMass

# Success Rate vs. Orbit Type

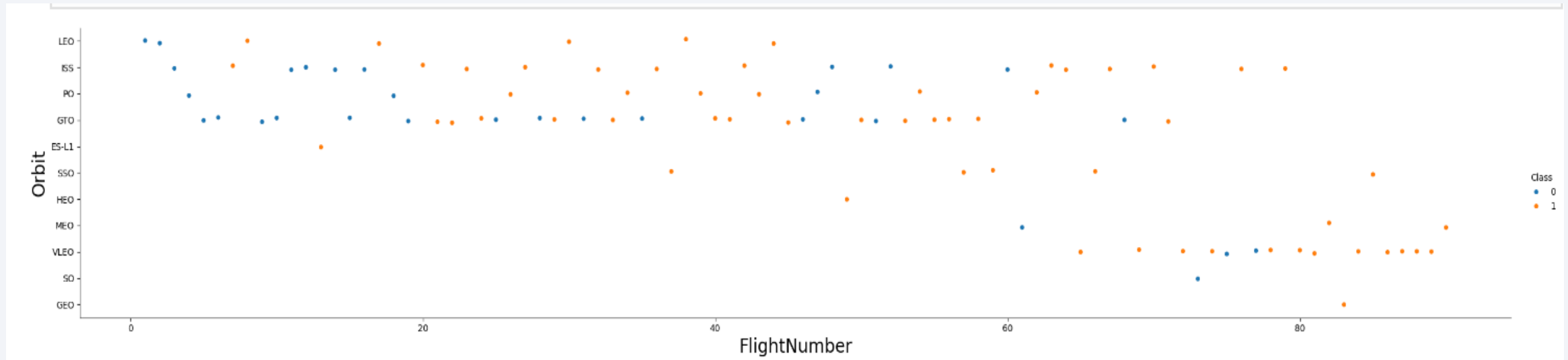
- Bar chart for the success rate of each orbit type
- ES-L1, GEO, HEO, SSO has the success rates while SO has 0% success rate followed by GTO.





# Flight Number vs. Orbit Type

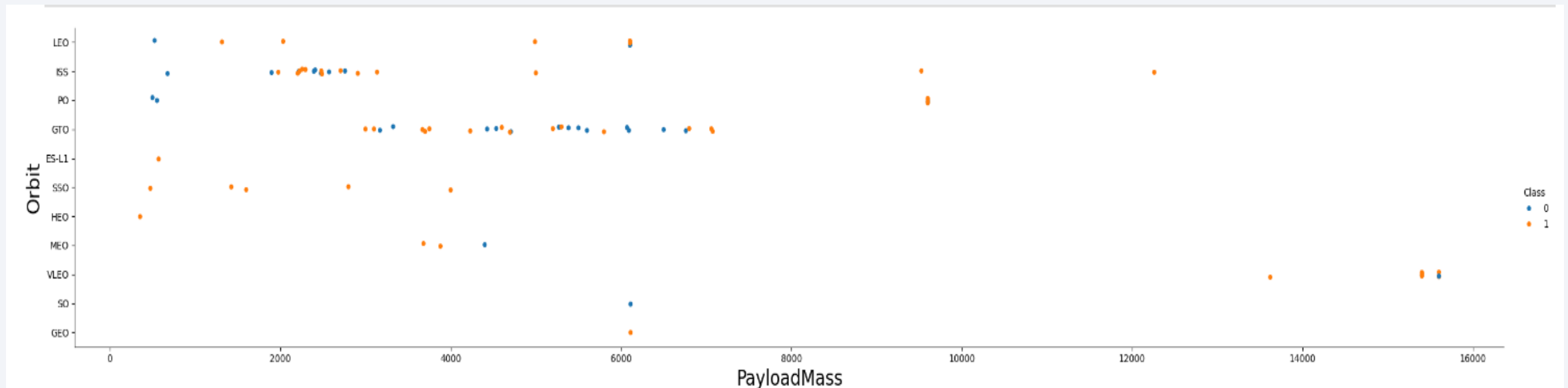
- Scatter point of Flight number vs. Orbit type



- Successful launches increase with FlightNumbers with VLEO having the most class 1 launches.

# Payload vs. Orbit Type

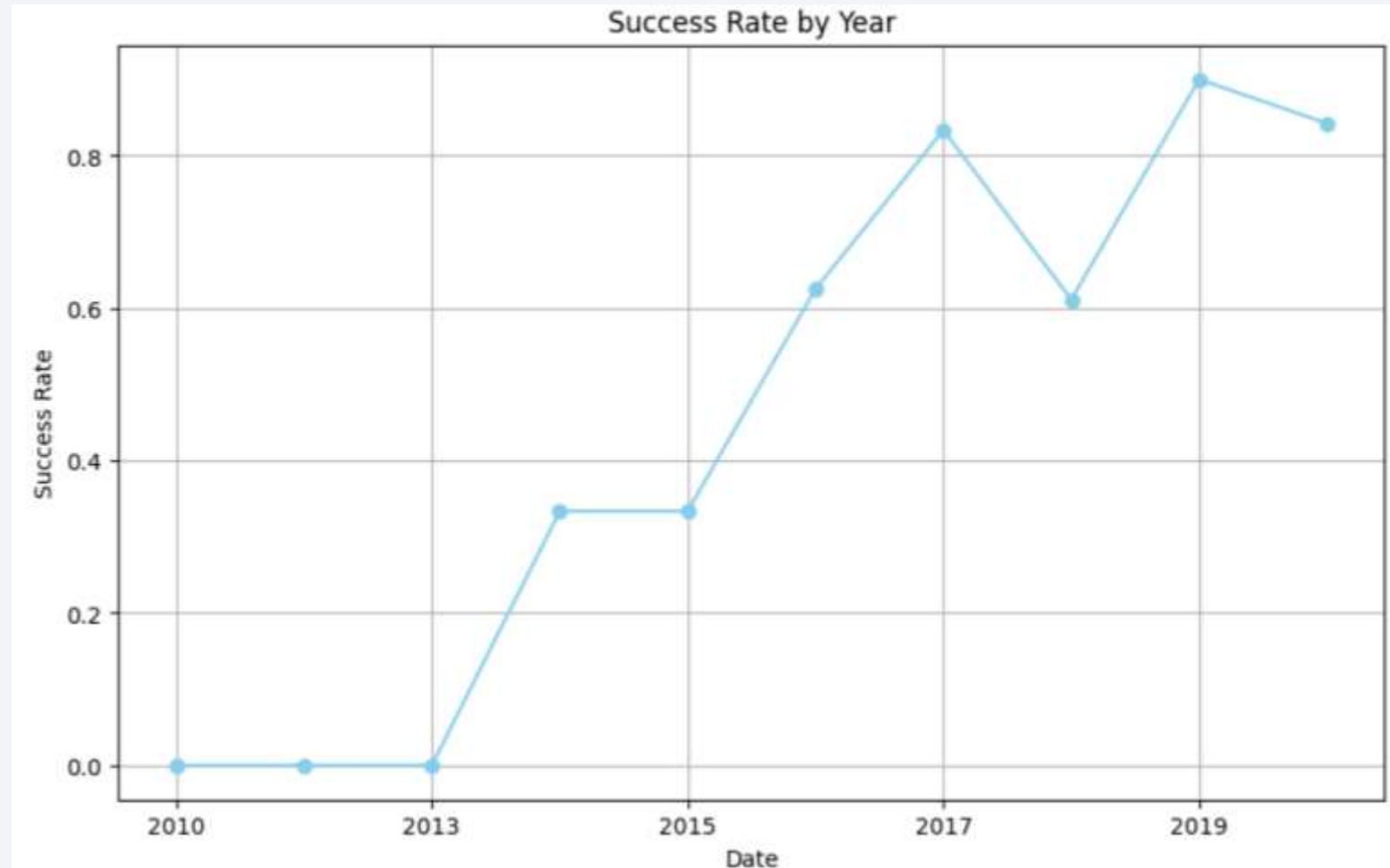
- Scatter point of payload vs. orbit type



- Success is more likely with GTO and ISS orbit types and lower payload mass launches are more likely to be successful in those orbits.

# Launch Success Yearly Trend

- Line chart of yearly average success rate
- Success rate increased steadily from 2013 until now except for a dip in 2017-18.
- However success rate is declining from 2019.



# 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

- There are 4 unique launch sites.

# Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- The adjoining query fetches the first 5 records that has the value from launch\_Site column starting with “CCA”.

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	S
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	S
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	S
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	S
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	S

# 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.
```

<b>sum(PAYLOAD_MASS__KG_)</b>
-------------------------------

45596
-------

- The total payload carried by NASA boosters is 45,596 KG.



# Average Payload Mass by F9 v1.1

---

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

```
%sql SELECT distinct(Booster_Version) FROM SPACEXTABLE;  
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

```
Done.
```

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

```
Done.
```

```
AVG(PAYLOAD_MASS__KG_)
```

```
2928.4
```

- The average payload carried by booster version F9v1.1 is 2982.4 KG.

# First Successful Ground Landing Date

---

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

```
%sql SELECT distinct(Landing_Outcome) FROM SPACEXTABLE;  
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

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

```
Done.
```

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

```
Done.
```

<b>MIN(Date)</b>
------------------

2015-12-22
------------

- First successful landing outcome on ground pad was 22-12-2015.

## 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 distinct(Landing_Outcome) FROM SPACEXTABLE;
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS_KG_ > 4000 AND PAYLOAD_MASS_KG_ < 6000 GROUP BY Booster_Version;
```

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

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

- There are 4 boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

# Total Number of Successful and Failure Mission Outcomes

---

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome,count(Mission_Outcome) FROM SPACEXTABLE group by Mission_Outcome;
```

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

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

- There 100 successful missions and 1 failed mission.

# Boosters Carried Maximum Payload

---

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

```
*sql SELECT Booster_Version,MaxPayload FROM (SELECT Booster_Version, MAX(PAYLOAD_MASS_KG_) as MaxPayload\
FROM SPACEXTABLE GROUP BY Booster_Version) AS subquery\
WHERE MaxPayload = (SELECT MAX(PAYLOAD_MASS_KG_) FROM SPACEXTABLE);
```

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

Booster_Version	MaxPayload
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

- The above are the boosters which carried the maximum payload of 15,600 KG.

# 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),Landing_Outcome, Booster_Version,launch_site\  
FROM SPACEXTABLE WHERE Landing_Outcome = "Failure (drone ship)"AND substr(Date,0,5)='2015'\  
OR Landing_Outcome = "Precluded (drone ship)" AND substr(Date,0,5)='2015';  
#%sql SELECT distinct(Landing_Outcome) FROM SPACEXTABLE;
```

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

Done.

substr(Date, 6,2)	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
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

- CCAFS LC-40 has 3 drone ship failed landing outcomes in 2015.



# 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
- The query fetches the landing outcomes between the dates provided by rank.

```
%sql SELECT Landing_Outcome, COUNT(*) as Outcome_Count,\nRANK() OVER (ORDER BY COUNT(*) DESC) as Rank\  
FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'\  
GROUP BY Landing_Outcome;
```

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

Landing_Outcome	Outcome_Count	Rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is dark blue with a thin white line representing the horizon. The city lights are visible as bright yellow and orange spots against the dark blue background of the Earth's surface.

Section 3

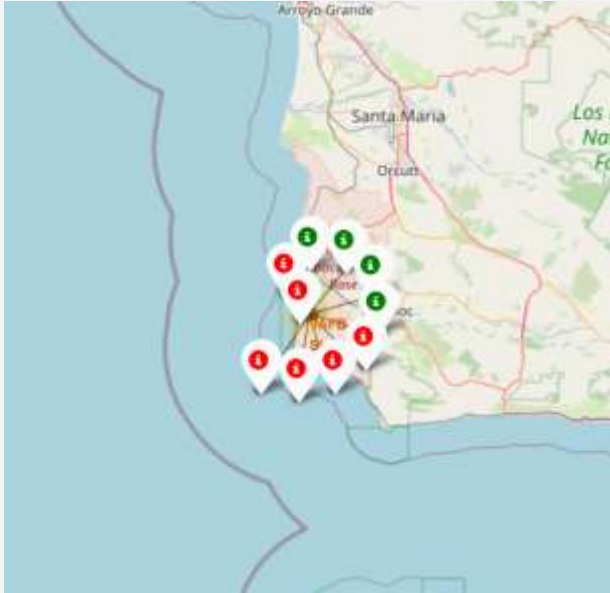
# Launch Sites Proximities Analysis

# Launch sites

- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- VAFB SLC-4E (California, USA)
  - Vandenberg Air Force Base Space Launch Complex 4E
- KSC LC-39A (Florida, USA)
  - Kennedy Space Center Launch Complex 39A
- CCAFS LC-40 (Florida, USA)
  - Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40 (Florida, USA)
  - Cape Canaveral Air Force Station Space Launch Complex 40



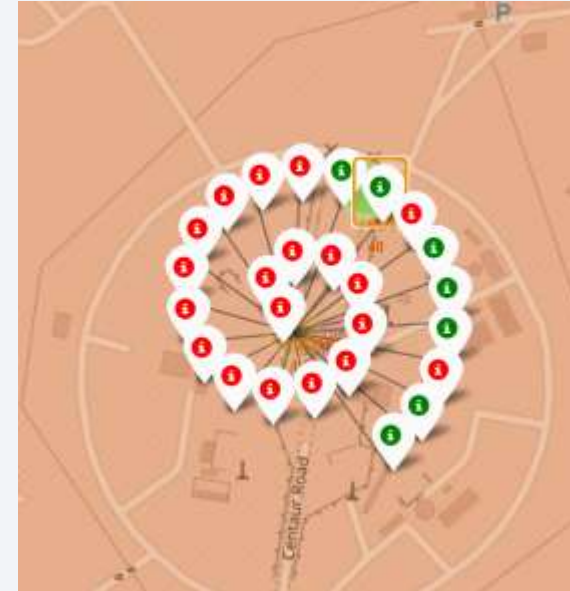
# Successful and failed launches by site



VAFB SLC-4E



KSC LC-39A



CCAFS LC-40



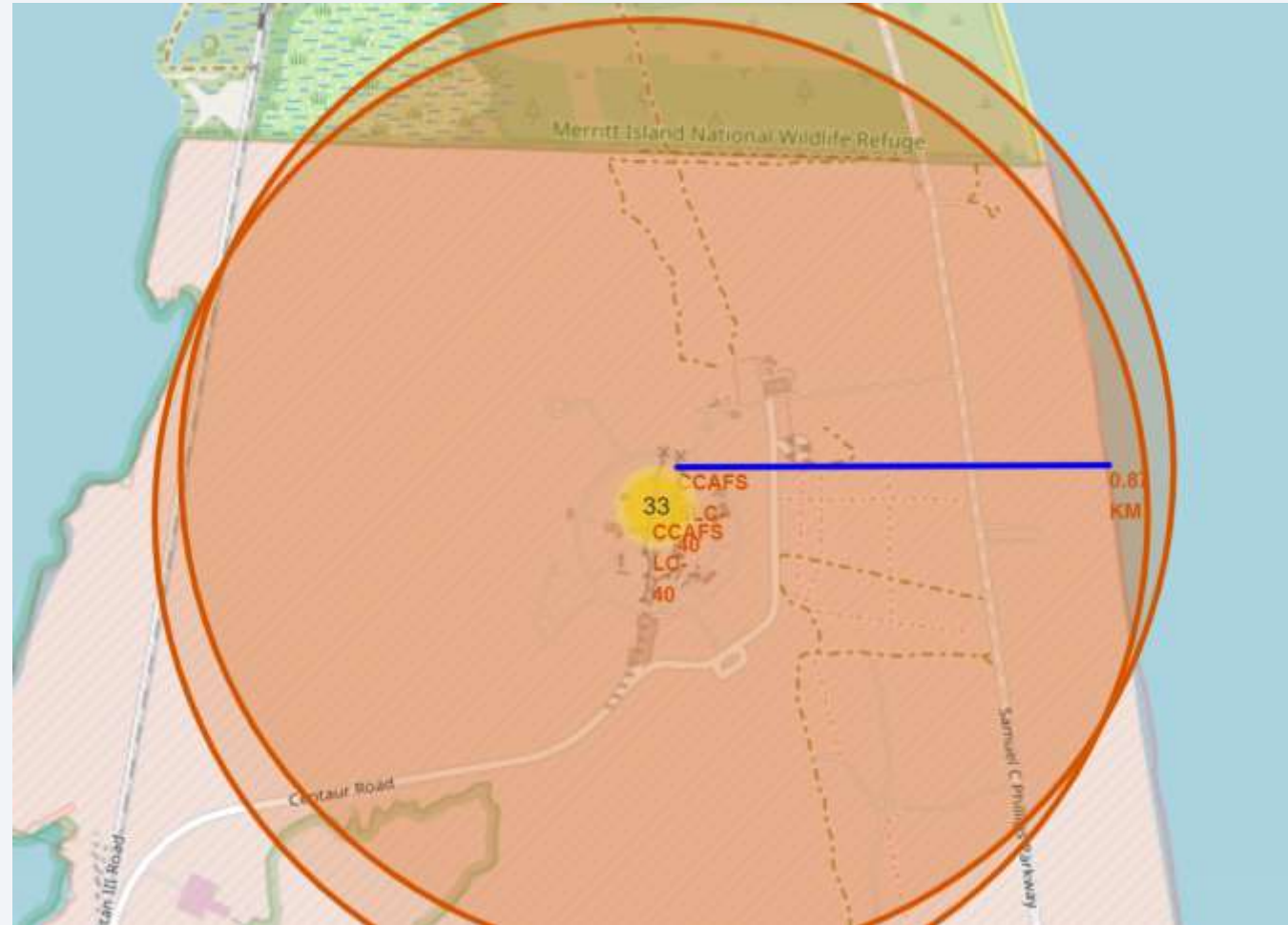
CCAFS SLC-40

- The Folium map shows the successful and failed launchings by site
- KSC LC-39A has the highest successful launchings followed by CCAFS LC-40.



# Distance to coastline

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- CCAFS SLC-40 has a distance of 0.87 Kms to nearest coastline.

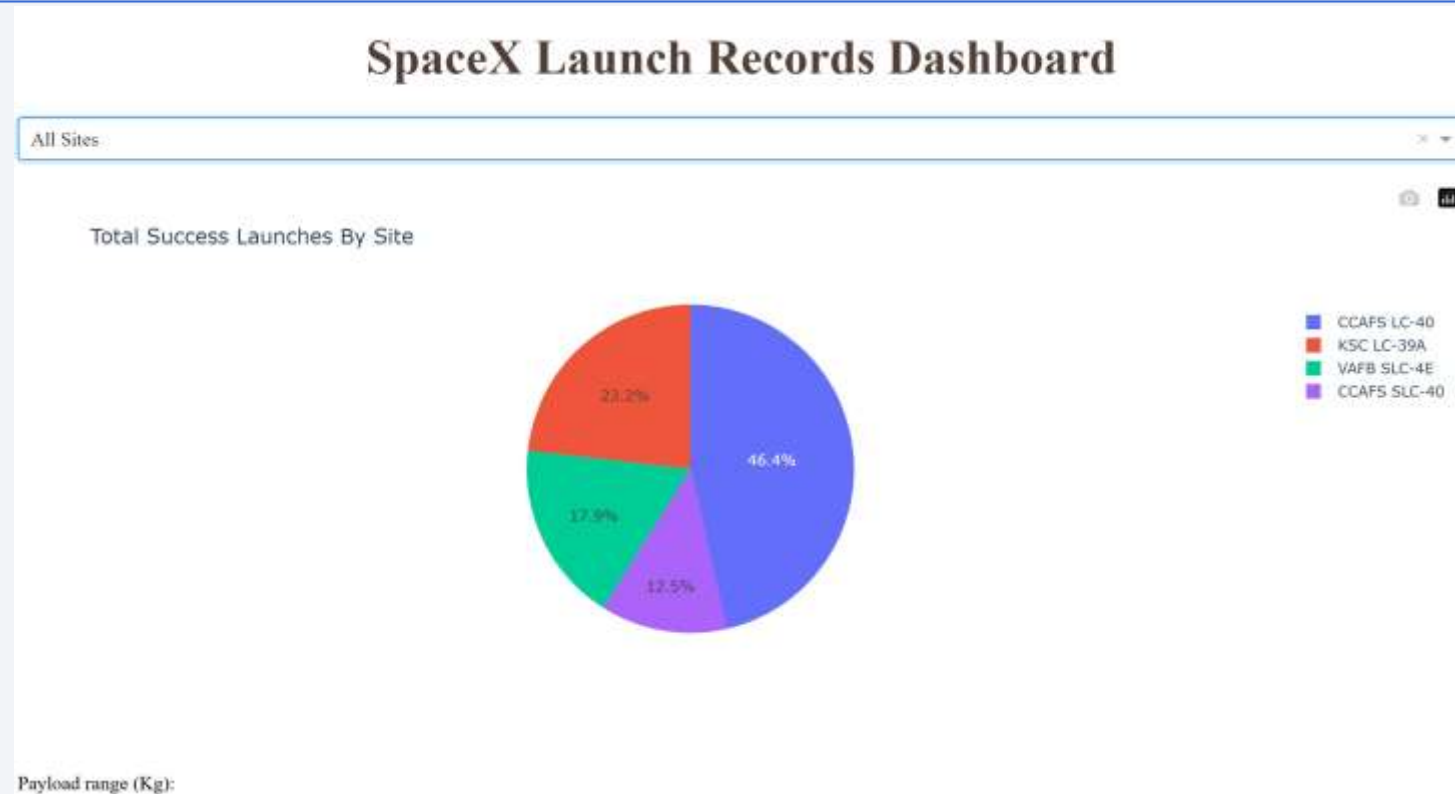




Section 4

# Build a Dashboard with Plotly Dash

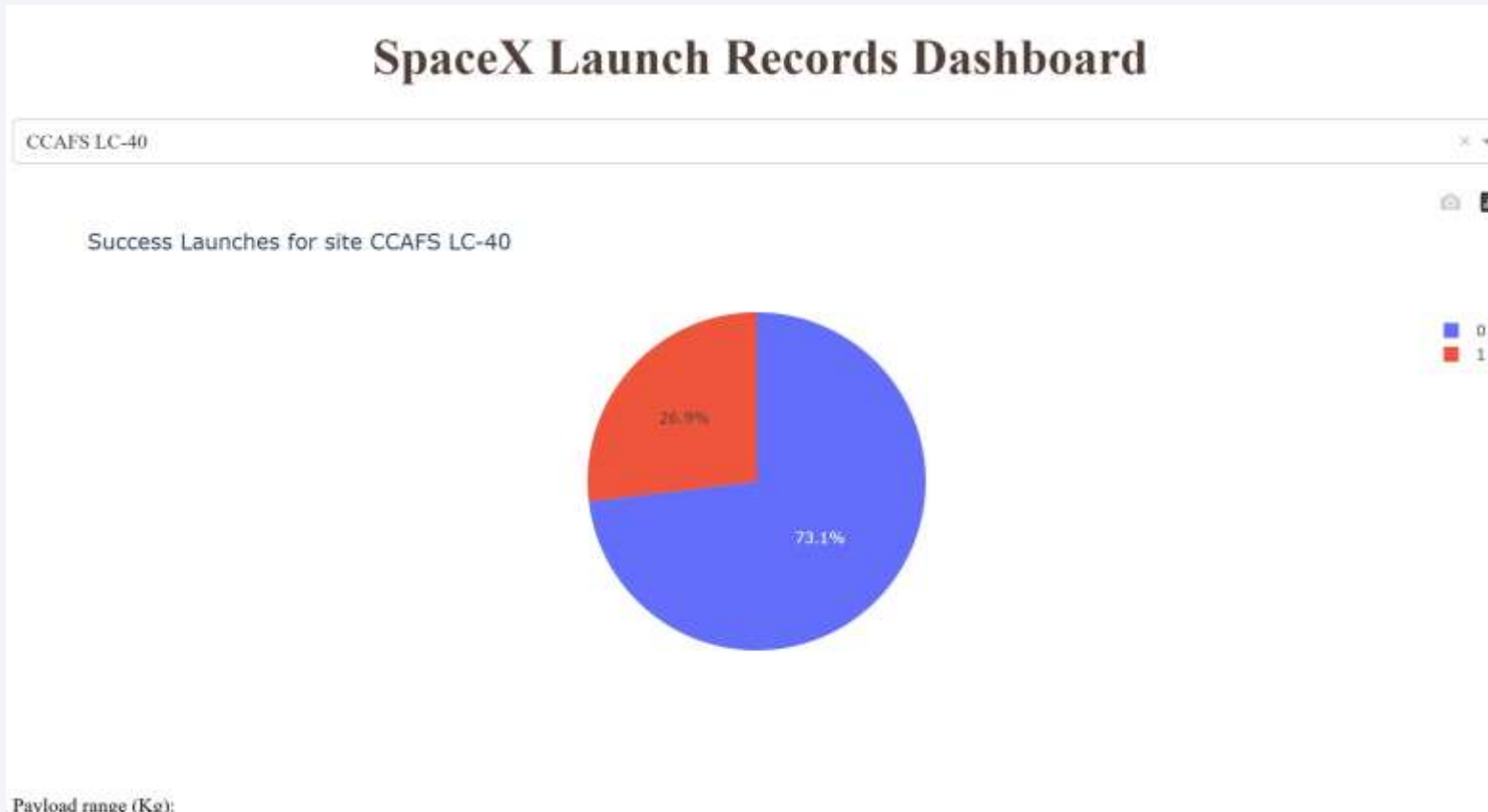
# Launch success count for all sites



- CCAFS LC-40 has highest percentage of launch success at 46.4% followed by KSC LC-39A at 23.2%.

# Launch site with highest launch success ratio

---



- CCAFS LC-40 has the highest launch success ratio at 26.9%



# Payload vs. Launch Outcome scatter plot for all sites



- FT Booster has highest success outcome across most payload ranges.



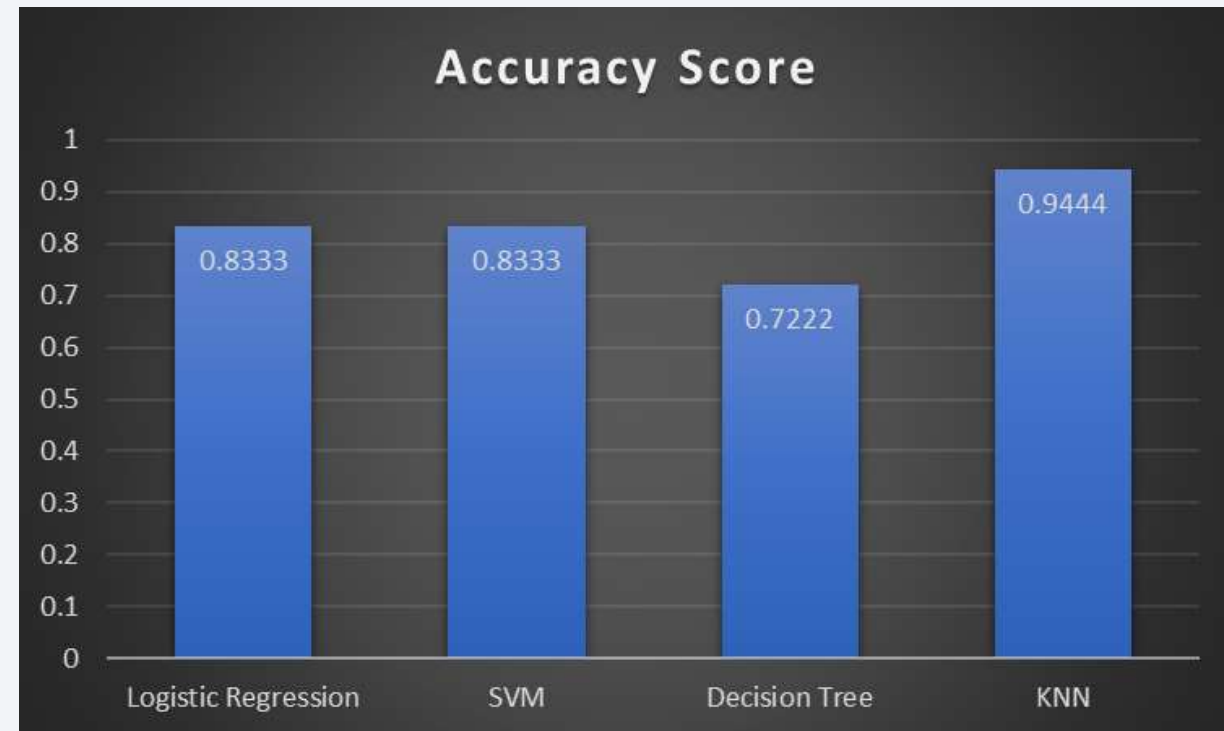
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

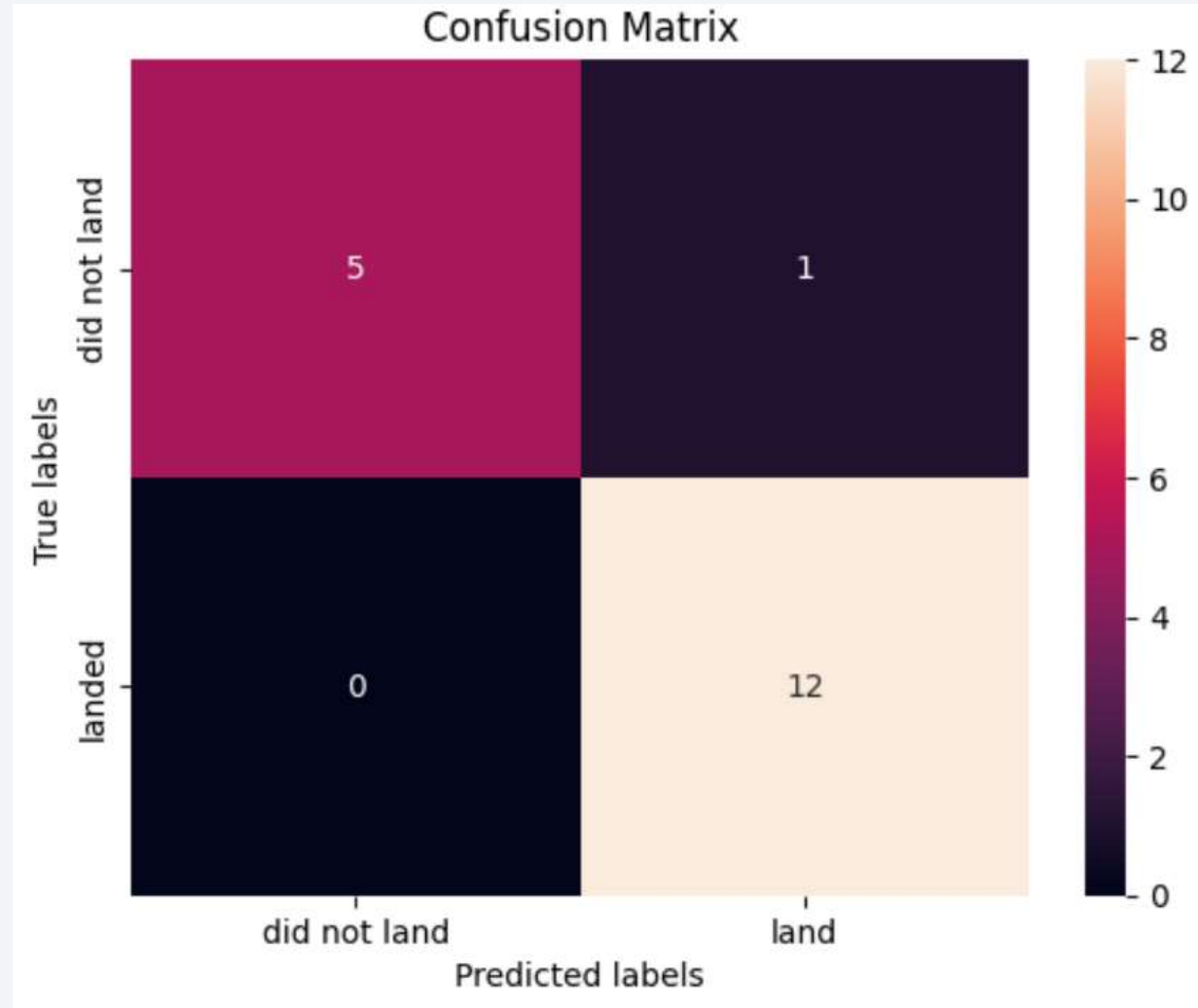
---

- Visualize the built model accuracy for all built classification models, in a bar chart
- K nearest neighbors had the best score at .9444 while Logistic Regression, SVM & KNN performed equally well at .8333 accuracy score on the test data. Decision tree had poor performance at .7222.



# Confusion Matrix

- K Nearest Neighbour was the best ML model with 5 true negatives, 1 false positive prediction, 0 false negatives and 12 true positives.



# Conclusions

---

- KSC LC-39A has the highest successful launchings followed by CCAFS LC-40.
- FT Booster has highest success outcome across most payload ranges.
- K Nearest Neighbour was the best ML model.
- Successful landings marked by Class 1 appears to increase as the flight numbers increase and CCAFS SLC 40 has most number of successful launches
- Successful landings increase with lower payload mass and with higher flight numbers

# Appendix

---

- [Mathewtmp/testrepo \(github.com\)](https://github.com/Mathewtmp/testrepo)

Thank you!

