IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Haider Sultan
30/11/22

# Outline

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

# Executive Summary

- **Summary of methodologies**:
  - Collection of Data
  - Data wrangling
  - Exploratory Data Analysis:
    1. SQL
    2. Data Visualization
  - Building interactive map with Folium
  - Building a Dashboard with Dash(Plotly)
  - Building Classification Models
- **Summary of all results**:
  - Exploratory data analysis results
  - Predictive analysis results

# Introduction

- **Project background and context:**

Space travel may soon be a possibility for common folks on this planet since the arrival of commercial space travel. Space X is currently leading this industry with prices as low as $62 million as compared to the $165 million charged by other providers.

This is mainly due to Space X's ability to recover part of rocket. If we can predict whether the first stage will land successfully, then we can determine the overall cost of the launch more effectively. Space Y can use this information to compete with SpaceX.

- **Problems you want to find answers:**

We want to figure out which variables affect out predicted outcomes the most and then determine the conditions to attain best results and ultimately predicting the success rate of Stage 1.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - We Combined data from SpaceX public API and used BeautifulSoup to scrape SpaceX Wikipedia page.
- Perform data wrangling
  - Generated Training labels by classifying landing as Successful/Unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used GridSearchCV to find the best Hyperparameters for our models.

# Data Collection

The data collection process includes a combination of API requests from the SpaceX API and web scraping data from a table in the Wikipedia page of SpaceX, Falcon 9 and Falcon Heavy Launches Records using BeautifulSoup.

**SpaceX API:**

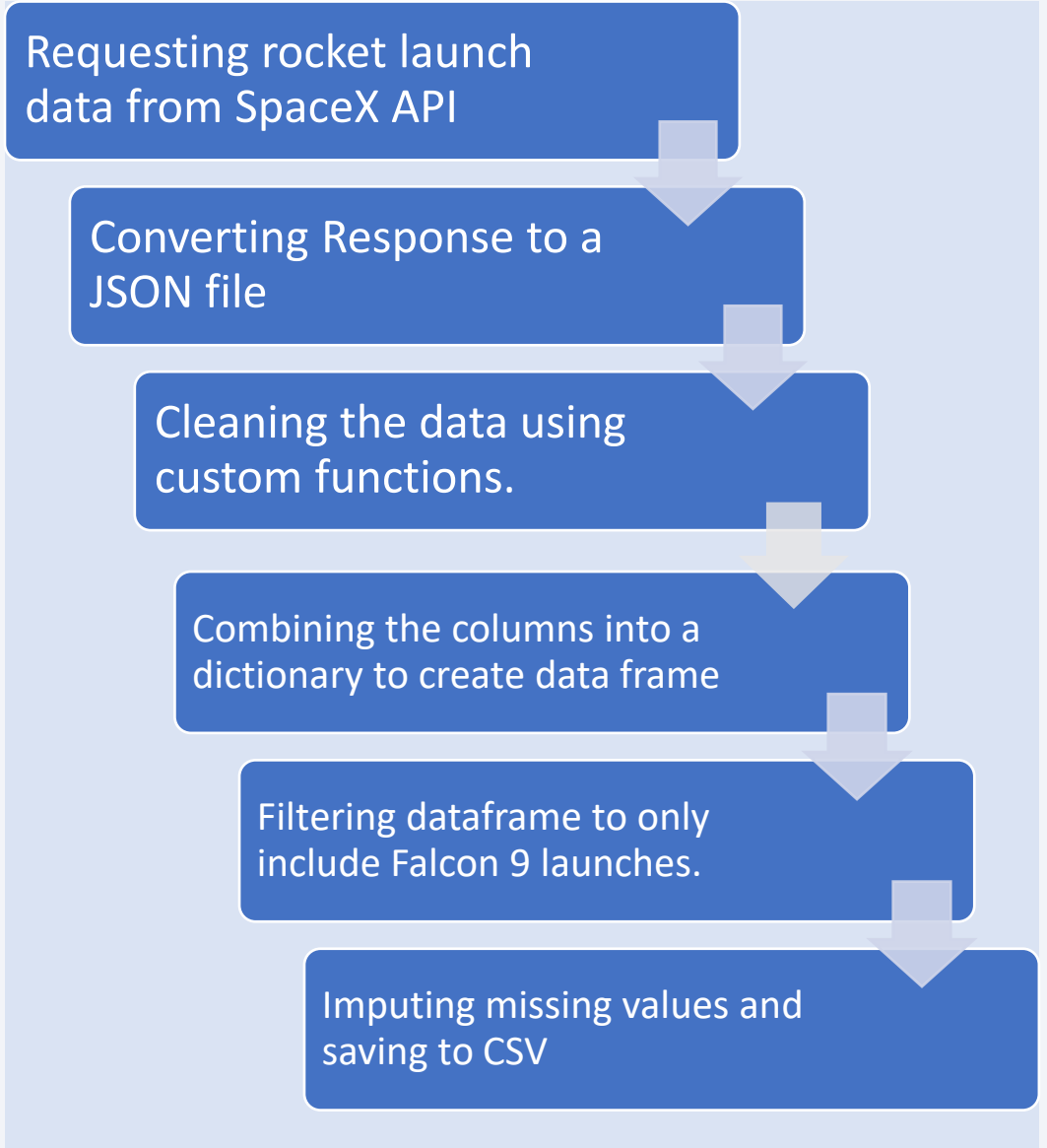
1. Request data from SpaceX API
2. Data is returned in the form of a .JSON
3. Data is Normalized and saved

**Web Scraping:**

1. Get HTML response from the URL
2. Extract data using BS4
3. Normalize and save as CSV

# Data Collection – SpaceX API

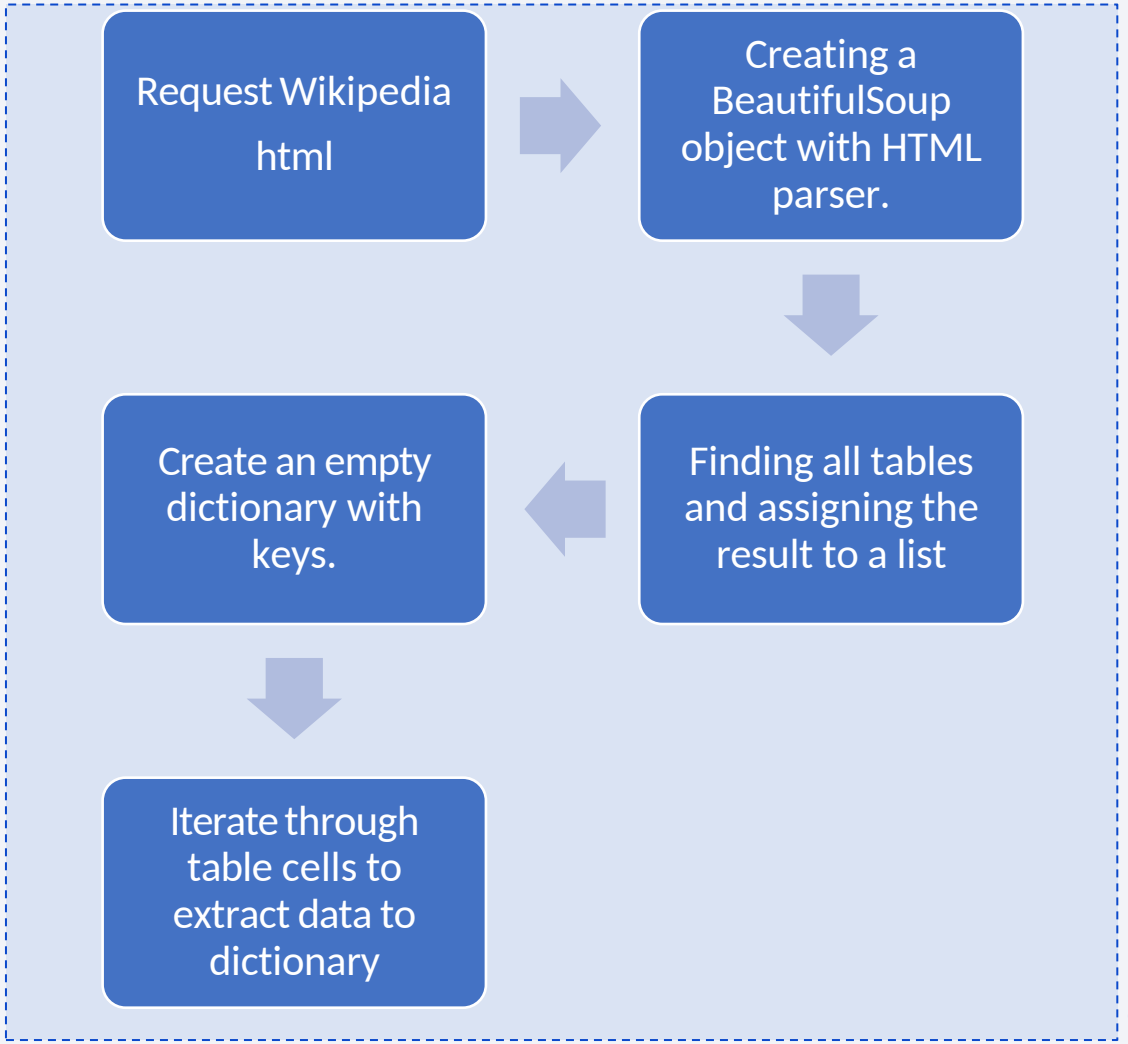
Github Link:
https://github.com/Hyde06/Capstone_SpaceX/blob/master/Capstone%20Project%20:%20SpaceX%20REST%20API.ipynb



# Data Collection - Scraping

- Github URL:

https://github.com/Hyde06/Capstone_SpaceX/blob/master/Capstone%20Project%20:%20Data%20Collection%20using%20bs4.ipynb



# Data Wrangling

- Create a training label with landing outcomes where successful = 1 & failure = 0.
- Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.
- Value Mapping:
    I. True ASDS, True RTLS, & True Ocean – set to -> 1
    II. None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0
- Calculating the success rate for every landing in dataset.

Github Link:
https://github.com/Hyde06/Capstone_SpaceX/blob/master/Project%20-%20Data%20Wrangling.ipynb

# EDA with Data Visualization

- Performed exploratory data analysis on dataset variables such as Flight number, Payload Mass, Orbit, Launch Site, Year and Class.
- Used Scatter plots, line charts and bar plots to determine any correlation between variables such as:
    1. Flight Number vs. Payload Mass,
    2. Flight Number vs. Launch Site,
    3. Payload Mass vs. Launch Site,
    4. Orbit vs. Success Rate,
    5. Flight Number vs. Orbit,
    6. Payload vs Orbit,
    7. Year vs Success Rate

Github URL:
https://github.com/Hyde06/Capstone_SpaceX/blob/master/EDA%20with%20Python%20Libraries.ipynb

# EDA with SQL

- Loaded data set into IBM DB2 Database and queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of  customers and booster versions, and landing outcomes.

- Github Link:

https://github.com/Hyde06/Capstone_SpaceX/blob/master/EDA%20with%20SQL.ipynb

# Build an Interactive Map with Folium

- Created objects and added to a folium map:
    1. Markers for all launch sites on a map
    2. Markers for the success/failed launches for each site on the map
    3. Lines showing the distances between a launch site to its proximities

- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

- Github Link:

https://github.com/Hyde06/Capstone_SpaceX/blob/master/Data%20Visualization%20-%20Folium.ipynb

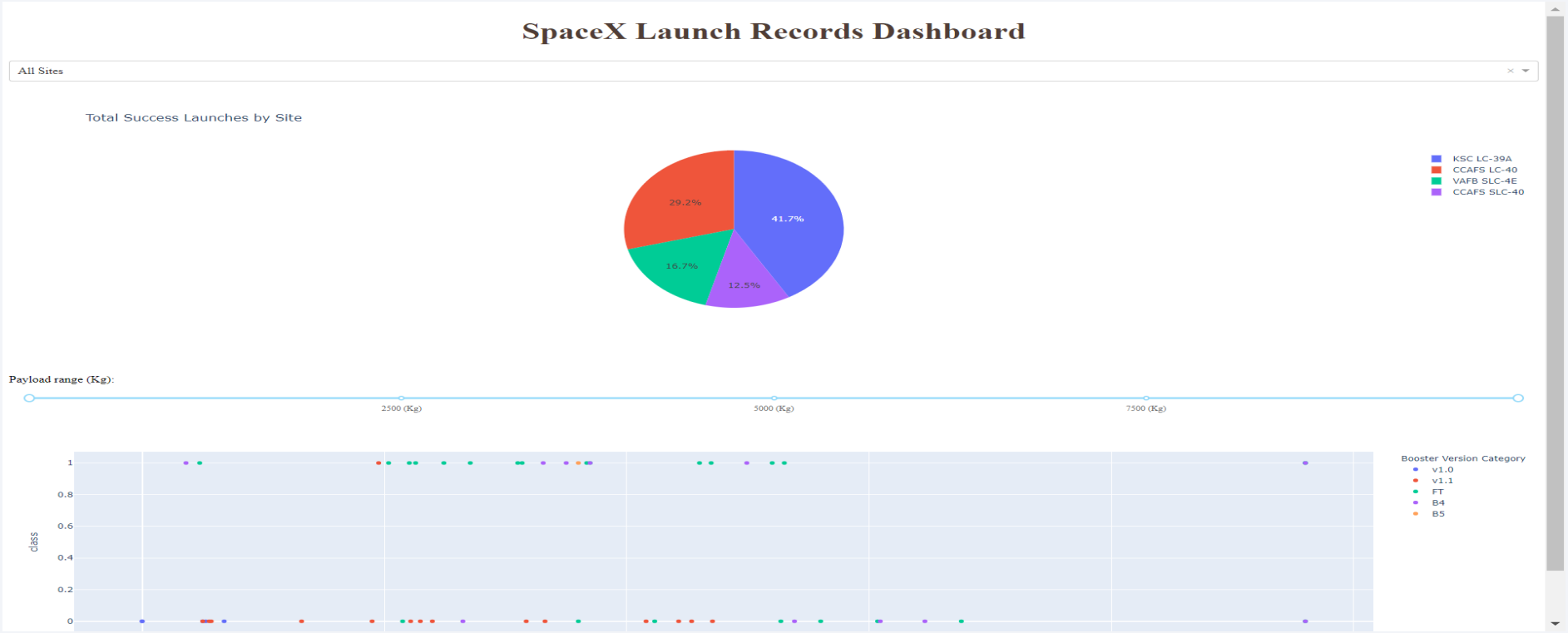
# Build a Dashboard with Plotly Dash

- Created Dashboards using Plotly Dash.
- It included a Pie chart and a Scatter Plot
- Pie chart can be selected to show distribution of successful landings across all launch sites.
- Scatter plot shows the relationship between Outcomes and Payload mass (Kg) by different boosters

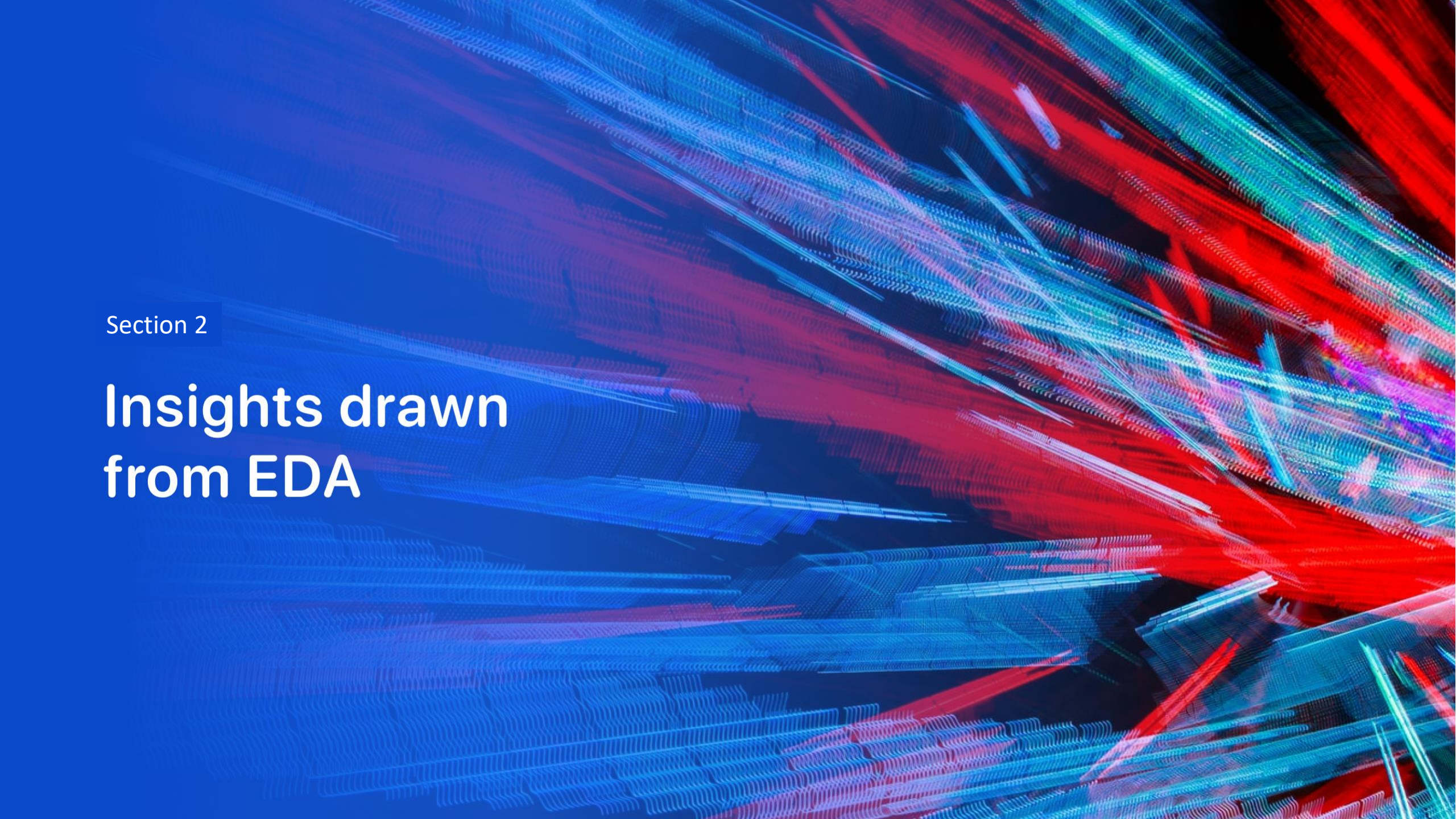
# Predictive Analysis (Classification)


Github URL:
https://github.com/Hyde06/Capstone_SpaceX/blob/master/Machine%20Learning%20Models.ipynb

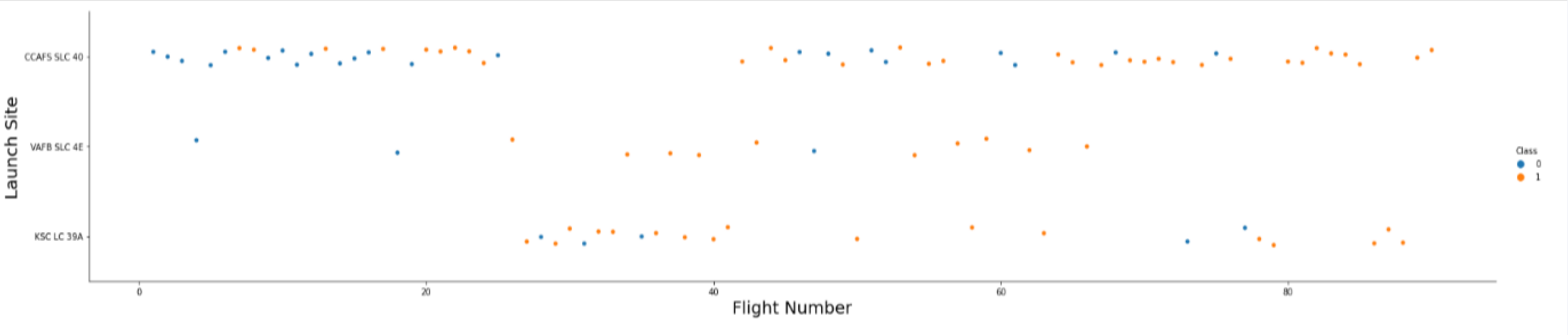
# Results




Plotly Dashboard is shown above. Comparing the accuracy of the four methods, all return the same accuracy of about *83%* for test data.

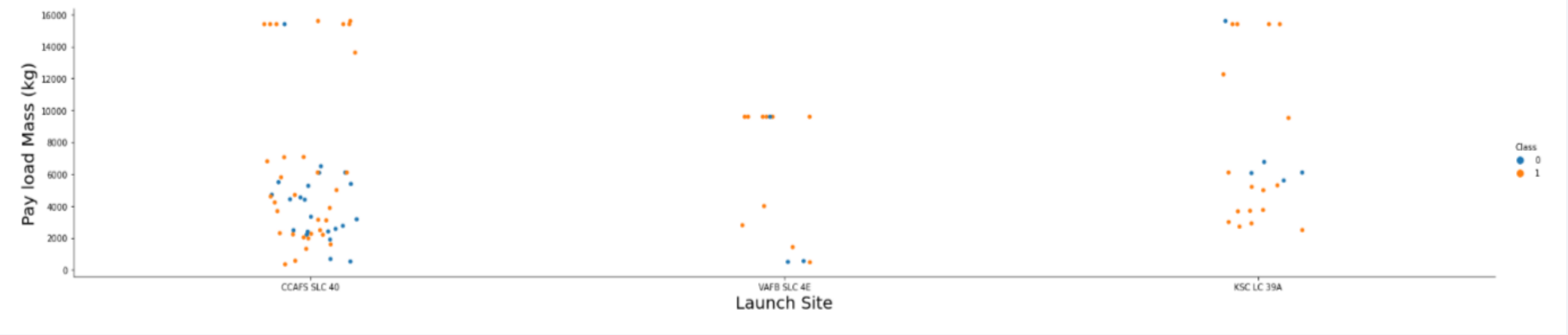

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site



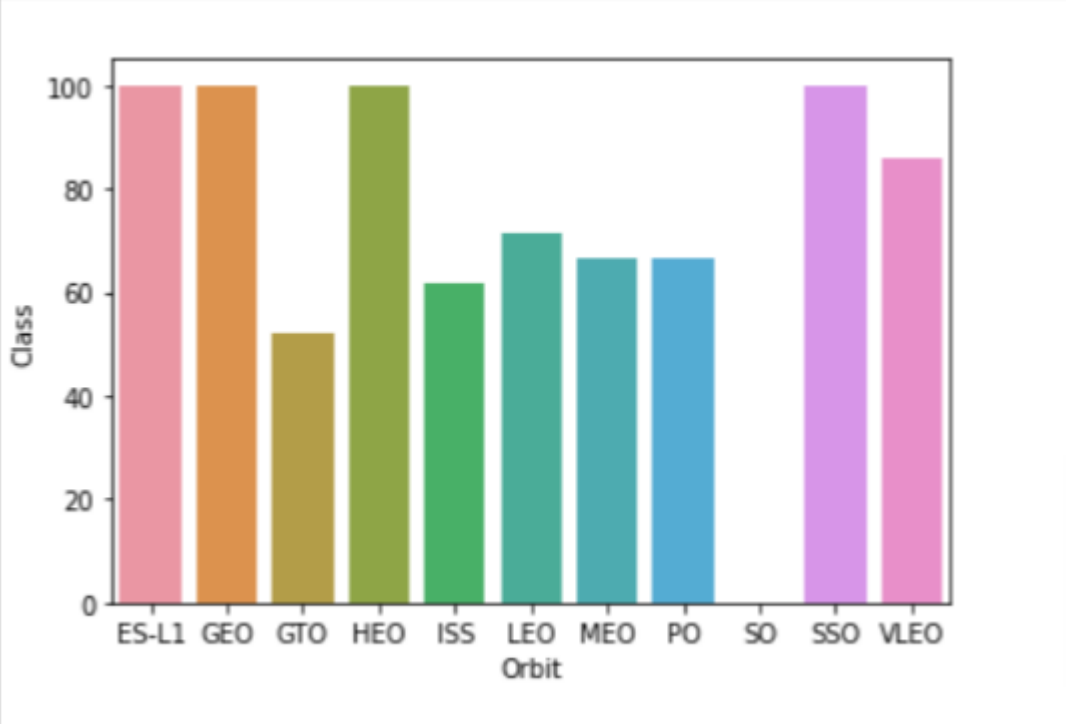

- Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

# Payload vs. Launch Site



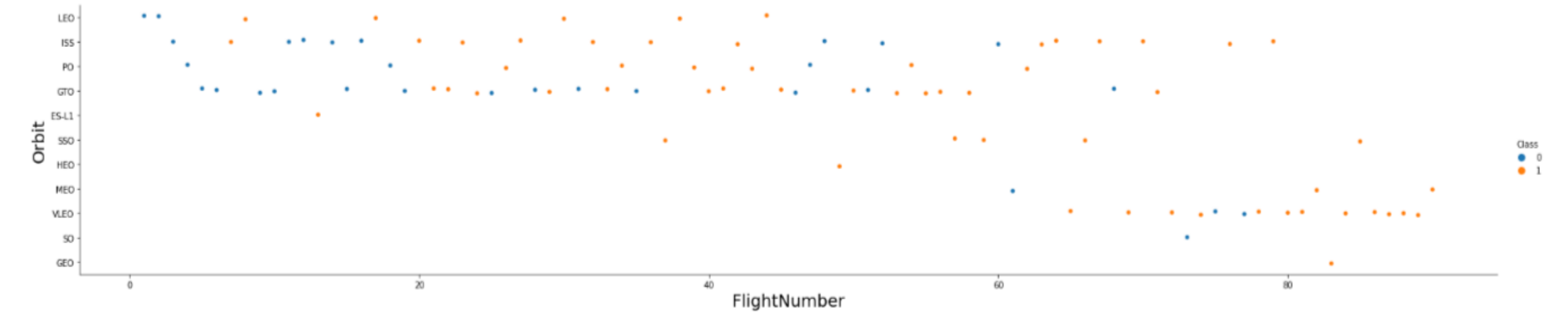

- In VAFB-SLC launch site, there are no rockets launched for heavy payload mass(greater than 10000).
- Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.

# Success Rate vs. Orbit Type



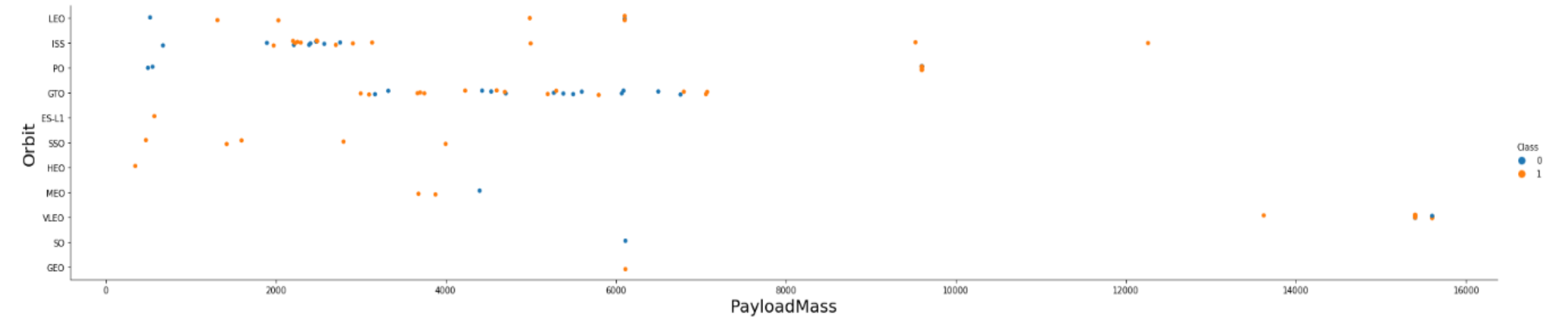

- Orbit types **SSO**, **HEO**, **GEO**, and **ES-L1** have **the highest success rates** (*100%*).
- GTO (27) has the around 50% success rate but largest sample.
- SO (1) has 0% success rate

# Flight Number vs. Orbit Type



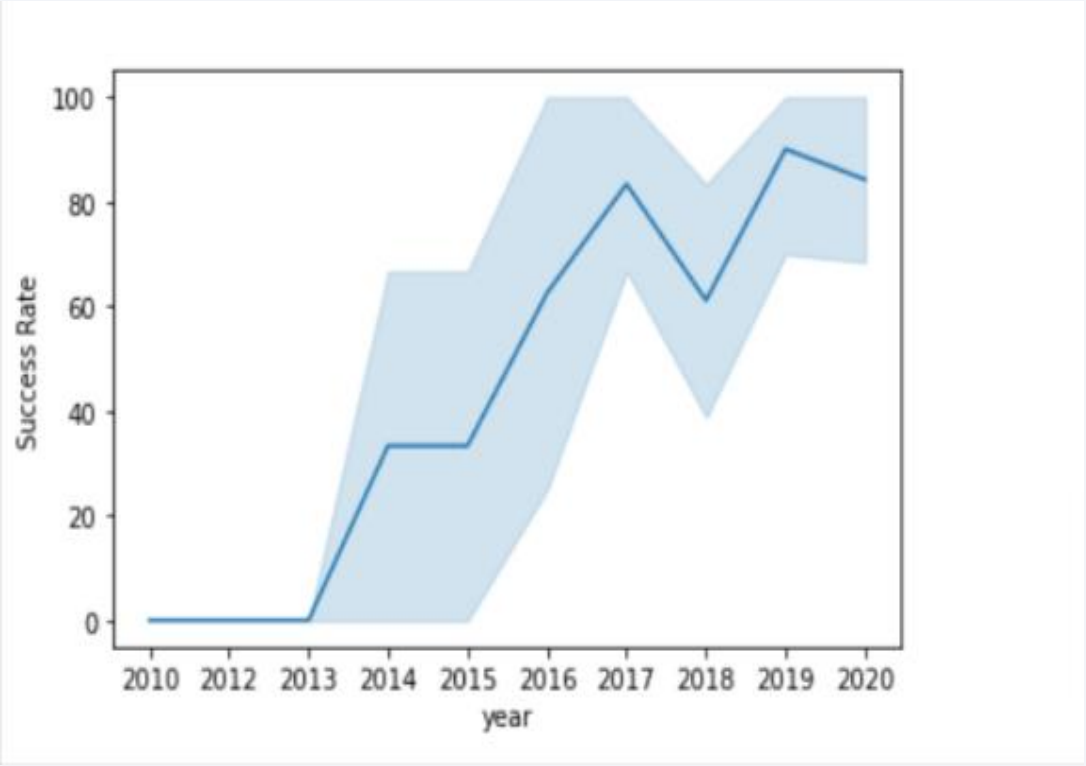

- Launch Orbit preferences changed over Flight Number.
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

# Payload vs. Orbit Type




- Payload mass seems to correlate with orbit
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

# Launch Success Yearly Trend




- Since 2013, the success rate has continued to **increase** until 2017.
- The rate decreased slightly in 2018.
- Recently, it has shown a success rate of about *80%*.
- The success rate since 2013 kept increasing till 2020.

# All Launch Site Names

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0
Done.
```

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

- Query unique launch site names from database.
- There are four unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

# Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

| 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 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 00:35:00 | 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 |

- Only five records of the SpaceX table were displayed using LIMIT 5 clause in the query.

# Total Payload Mass

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databa
Done.
```

| SUM |
| --- |
| 45596 |

- Using the SUM() function to calculate the sum of column PAYLOAD_MASS__KG_.

# Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdoma
Done.
```

| average |
| --- |
| 2534 |

- Using the AVG() function to calculate the average value of column PAYLOAD_MASS__KG_ with a WHERE clause as well.

# First Successful Ground Landing Date

```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde
Done.
```

| DATE |
| --- |
| 2010-06-04 |

- This query returns the first successful ground pad landing date.
- Using the MIN() function to find out the earliest date in the column DATE.

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTBL where (mission_outcome like 'Success') AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

| booster_version |
| --- |
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

- This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000.

# Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

| mission_outcome | COUNT |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

- Using the COUNT() function to calculate the total number of columns.
- According to the result, *SpaceX* seems to have successfully completed 99 of its missions.

# Boosters Carried Maximum Payload

```
maximum = %sql select max(payload_mass__kg_) from SPACEXTBL
max2 = maximum[0][0]
%sql select booster_version from SPACEXTBL
where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTBL)
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0l
Done.
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0l
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 |

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

# 2015 Launch Records

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site
from SPACEXTBL where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.data
Done.
```

| MONTH | landing__outcome | booster_version | launch_site |
|---|---|---|---|
| January | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| April | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

- This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.
- Returned 2 such occurrences.

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

```
%sql select landing__outcome, count(*) as count
from SPACEXTBL where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing__outcome ORDER BY count Desc
```

```
 * ibm_db_sa://sng32430:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj
Done.
```

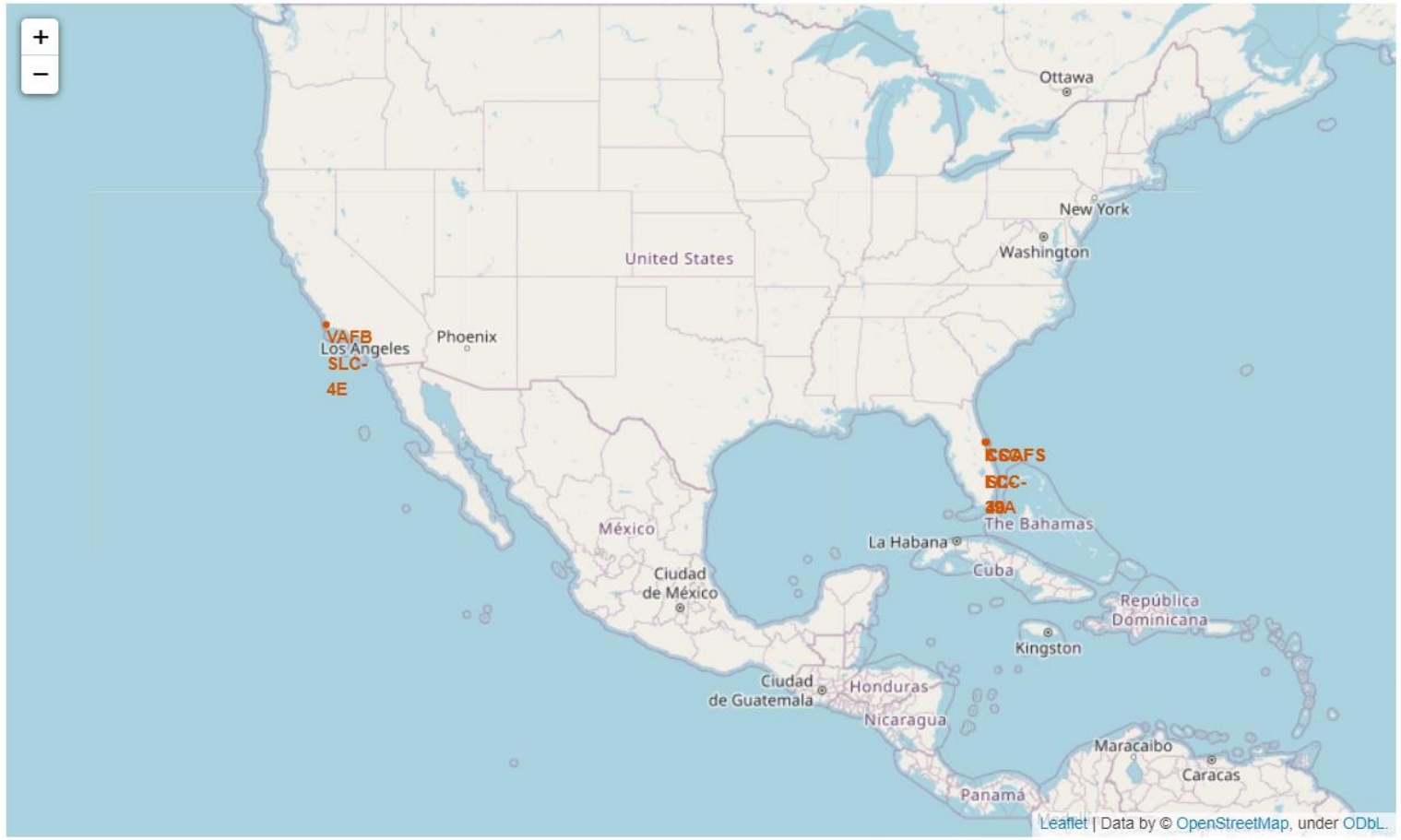
| landing__outcome | COUNT |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

- This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.
- According to the results, the number of successes and failures between 2010-06-04 and 2017-03-20 was similar.

Section 3

# Launch Sites Proximities Analysis

# <Folium Map Screenshot 1>

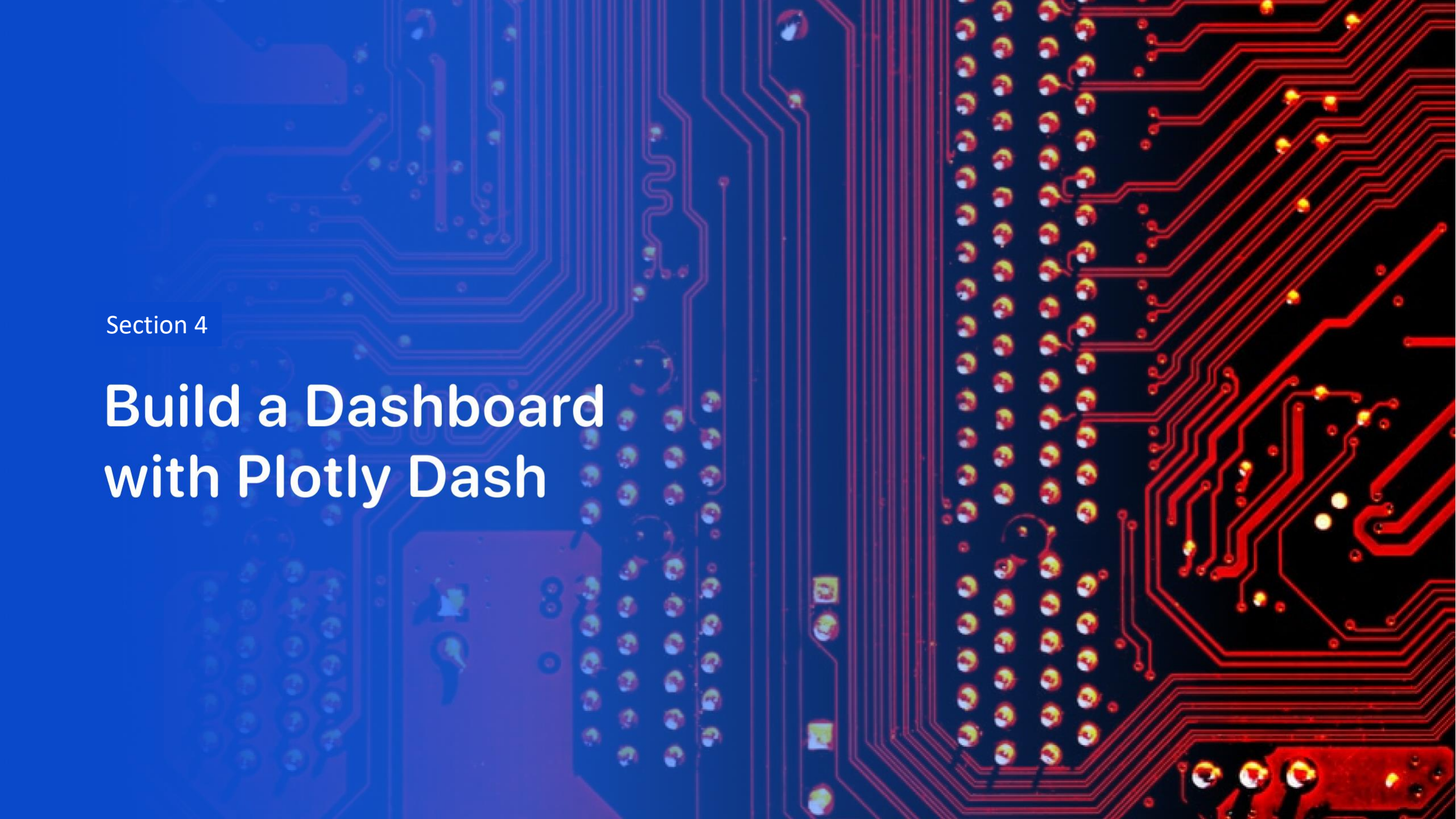




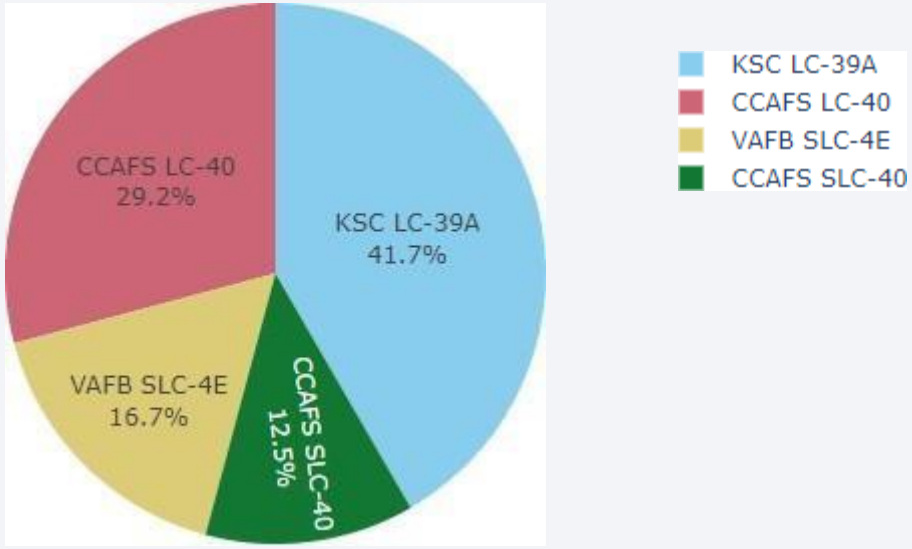
- The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

# <Folium Map Screenshot 2>

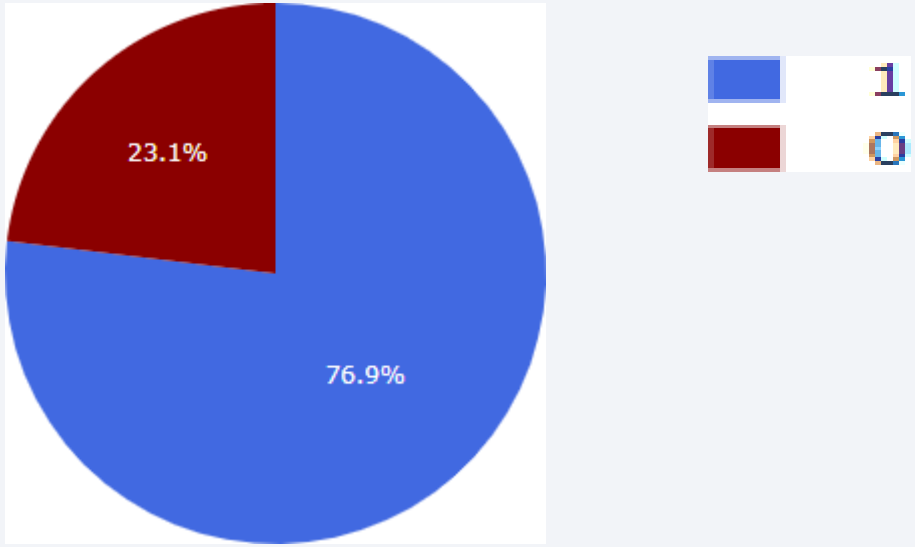

- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.


Section 4
Build a Dashboard with Plotly Dash

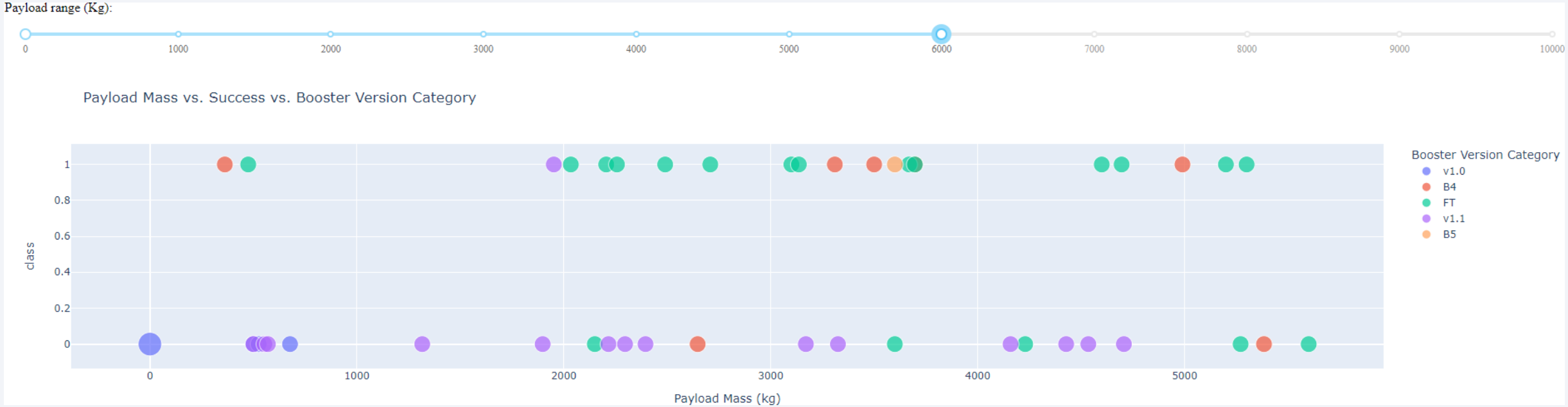
# Successful Launches Across Launch Sites



# Highest Success Rate Launch Site



# Payload Mass vs. Success vs. Booster  Version Category



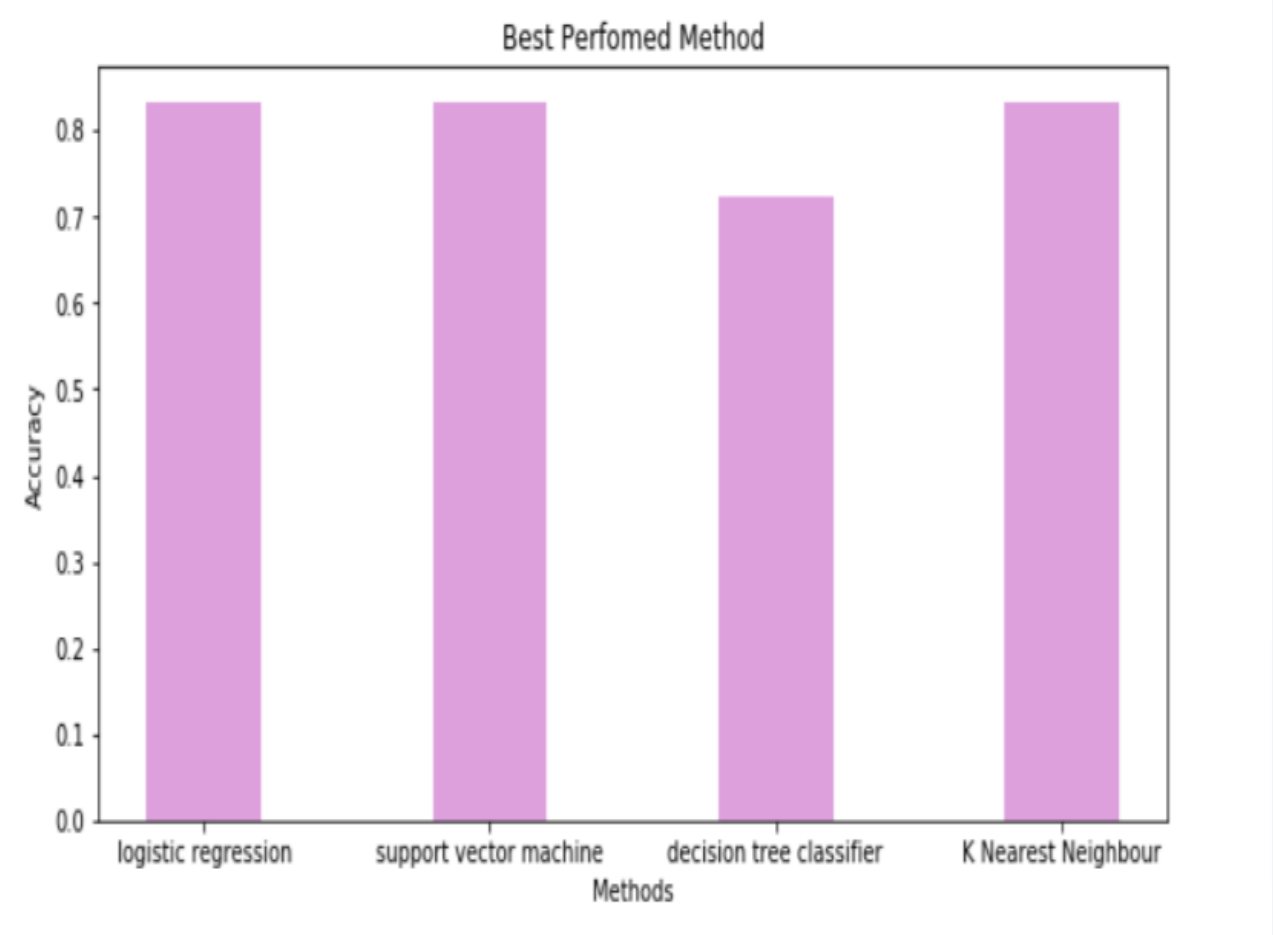
Section 5

# Predictive Analysis (Classification)

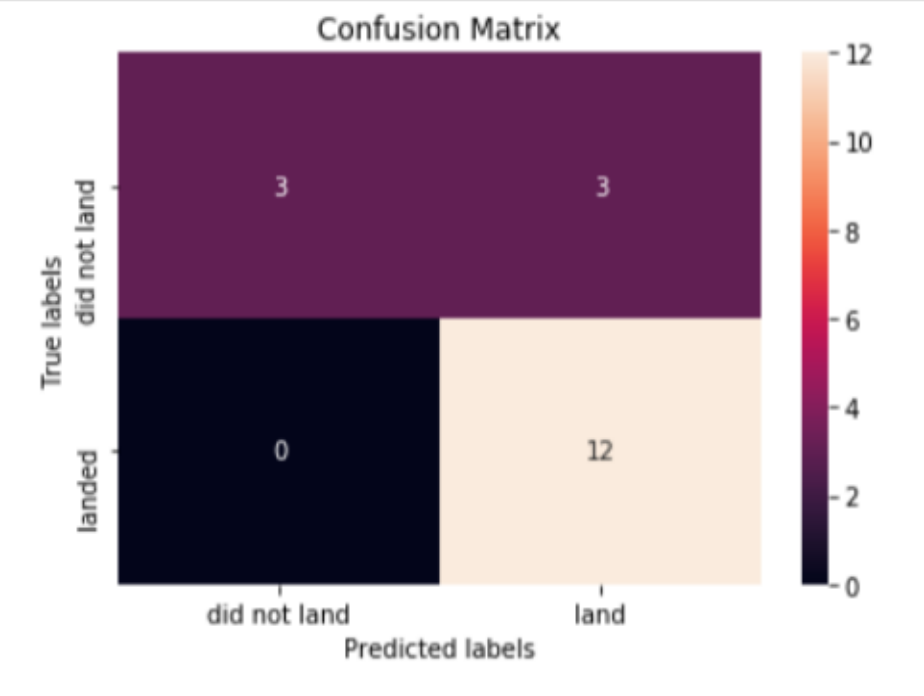
# Classification Accuracy

| | |
|---|---|
| **0.833333** | logistic regression |
| **0.833333** | support vector machine |
| **0.722222** | decision tree classifier |
| **0.833333** | K Nearest Neighbour |

- In the test set, the accuracy of all models was virtually the same at 83.33%.(Decision tree gives varying accuracy)
- It should be noted that the test size was small at 18.
- Therefore, more data is needed to determine the optimal model.



# Confusion Matrix




- Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.

# Conclusions

- As the number of flights increased, the success rate increased, and recently it has exceeded 80%.
- Orbital types SSO, HEO, GEO, and ES-L1 have the highest success rate (100%).
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- In this dataset, all models have the same accuracy (83.33%), but it seems that more data is needed to determine the optimal model due to the small data size.
- KSLC-39A has the highest number of launch successes and the highest success rate among all sites.
- The launch success rate of low weighted payloads is higher than that of heavy weighted payloads.

# Appendix

- Github URL:

https://github.com/Hyde06/Capstone_SpaceX

- Coursera Applied Data Science Capstone Course URL:

https://www.coursera.org/professional-certificates/ibm-data-science?

# Thank you!