



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Simge Karabasak  
6 July 2025



# Outline

---

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

# Executive Summary

---

- Defining the problem and the objectives
- Methodology
  - Data gathering
  - Data Analysis
  - Data Visualization
- Analysing the results
- Discussion on findings and limitations
- Final conclusion and future steps

# Introduction

---

- Our biggest competitor, Space X offers the most affordable rocket launches in the sector
- SpaceX differs from its competitors because it can recover the first stage and reuse it again
- The first stage is quite large and expensive. Besides most of the work is done by the first stage
- There are different factors that have effect on reusability of the first stage
- Understanding these factors in order to compete with SpaceX

Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - Web Scraping, APIs (SpaceX REST API), Request library
- Perform data wrangling
  - Sampling data
  - Dealing with missing values and outliers
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Train different classification models
  - Optimise the Hyperparameter grid search
  - Build a predictive model in order to function more efficiently

# Data Collection

---

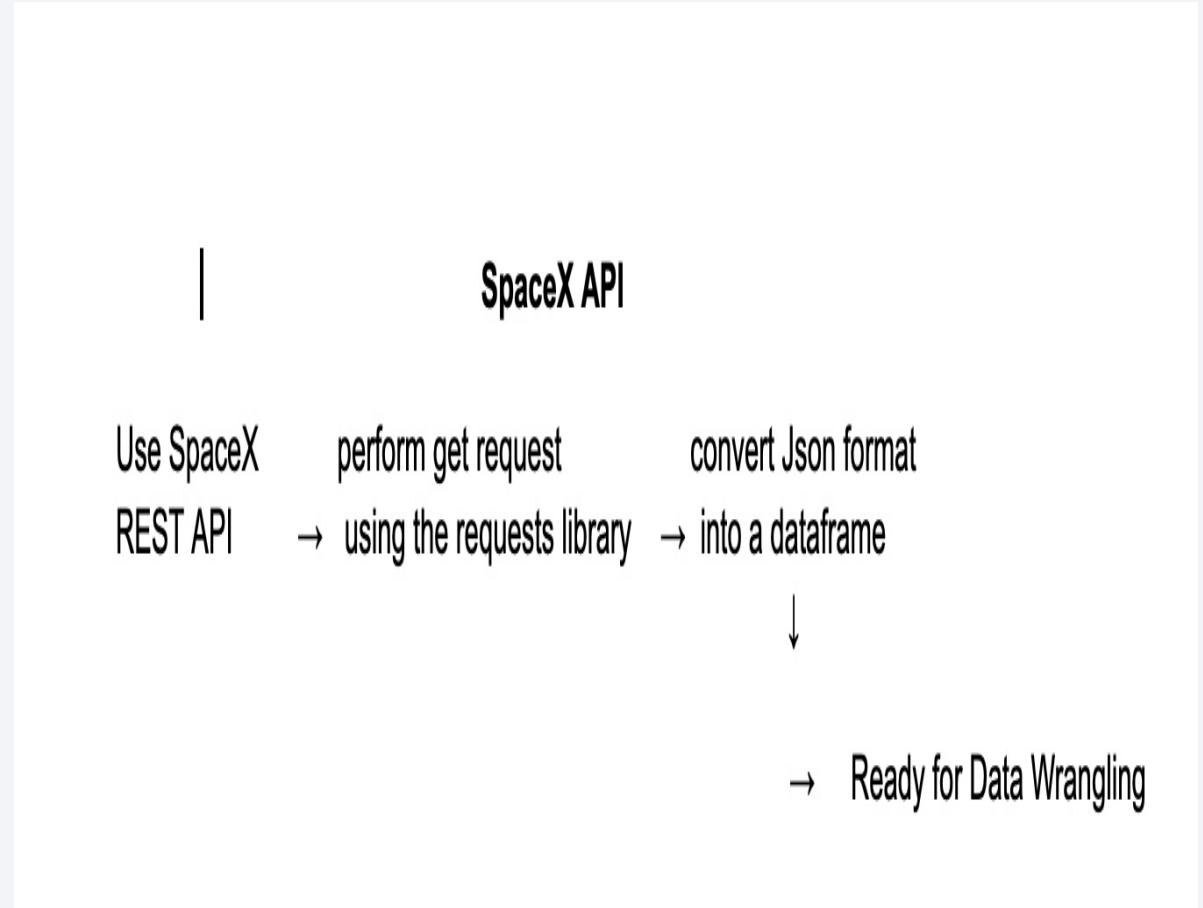
- SpaceX launch data is gathered from SpaceX REST API which provides data about launches, the rocket used, payload delivered, launch specifications, landing specifications and landing outcome.
- Perform a *get request* using the requests library to obtain the launch data, which used to get the data from the API.
- The json\_normalize function were used to convert the JSON format to a dataframe
- Another data source for Falcon 9 Launch Data is web scrape related Wiki pages with the help of BeautifulSoup package
- Parse the data from tables and convert them into a Pandas data frame

# Data Collection – SpaceX API

---

- Flowchart of SpaceX API

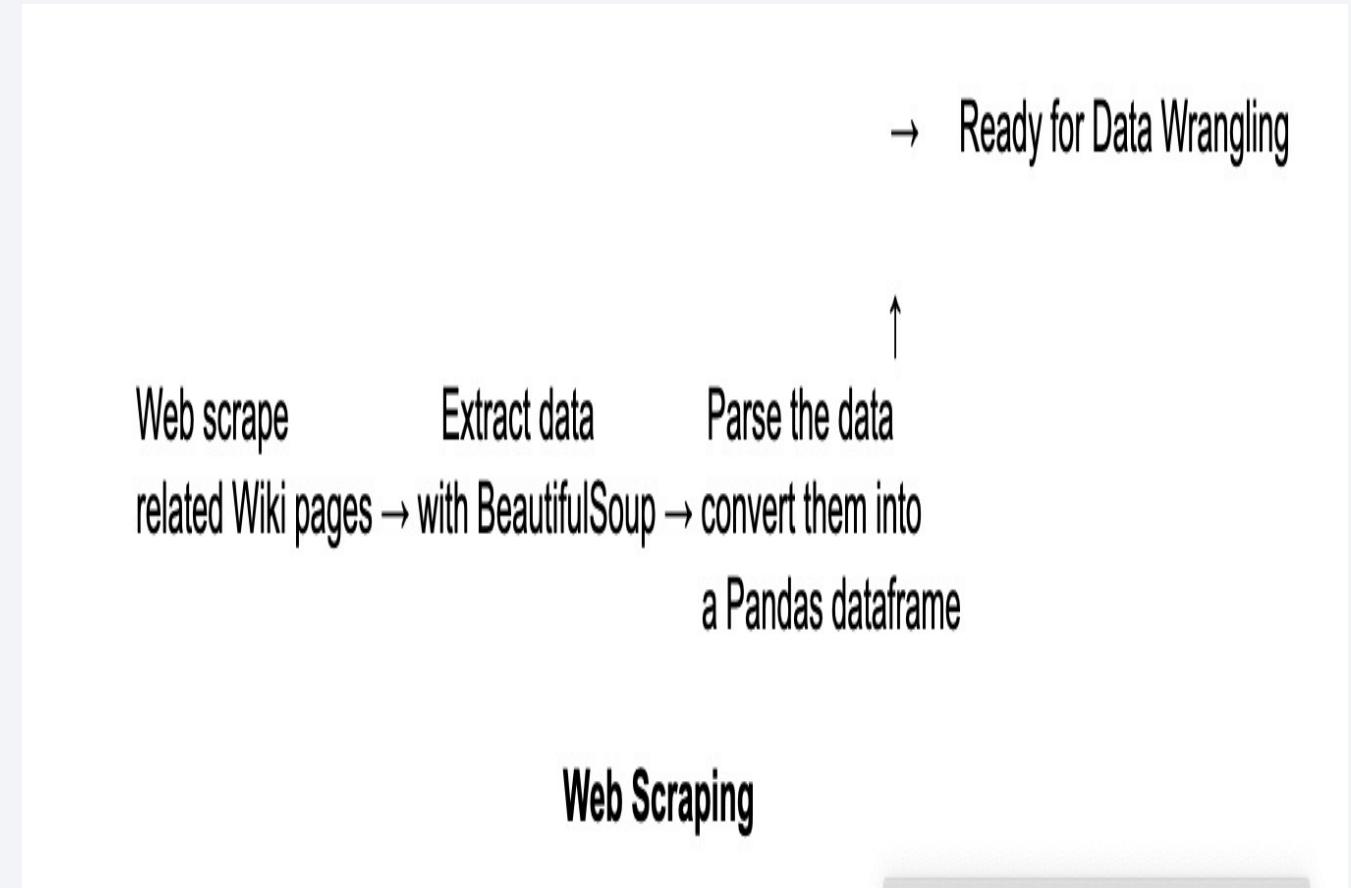
[https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX\\_Data\\_Lab/spacex-data-collection-api.ipynb](https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX_Data_Lab/spacex-data-collection-api.ipynb)



# Data Collection - Scraping

- The flowchart of Web Scraping

[https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX\\_Data\\_Lab/Webscraping.ipynb](https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX_Data_Lab/Webscraping.ipynb)



# Data Wrangling

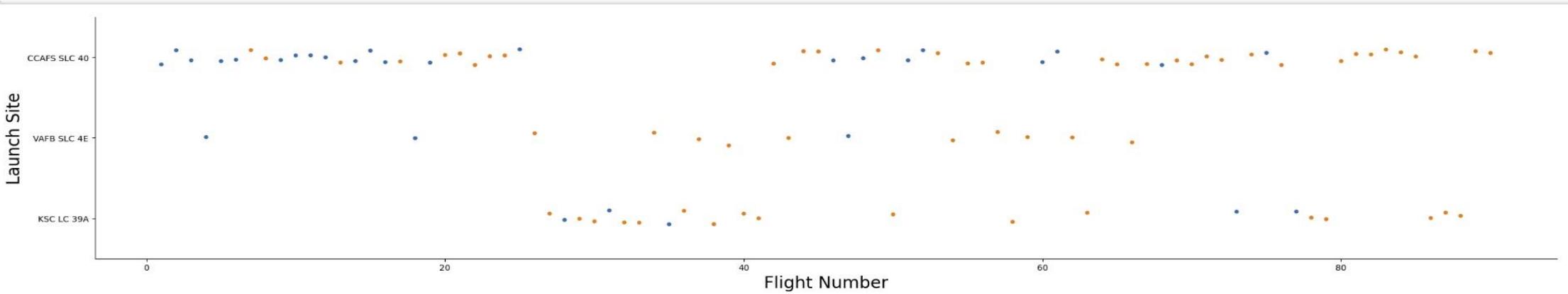
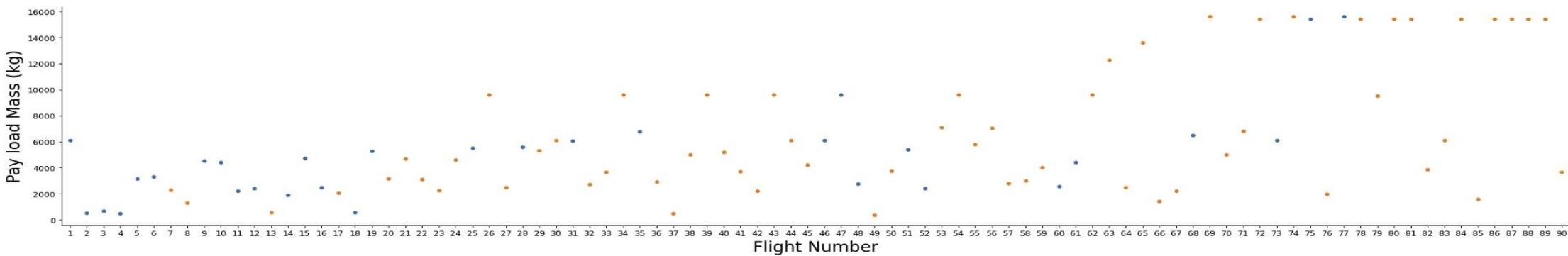
---

- Remove blank rows from table
- List the attributes in order to understand the data
- Rank the outcomes to determine relationship

[https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX\\_Data\\_Lab/labs-jupyter-spacex-Data%20wrangling.ipynb](https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX_Data_Lab/labs-jupyter-spacex-Data%20wrangling.ipynb)

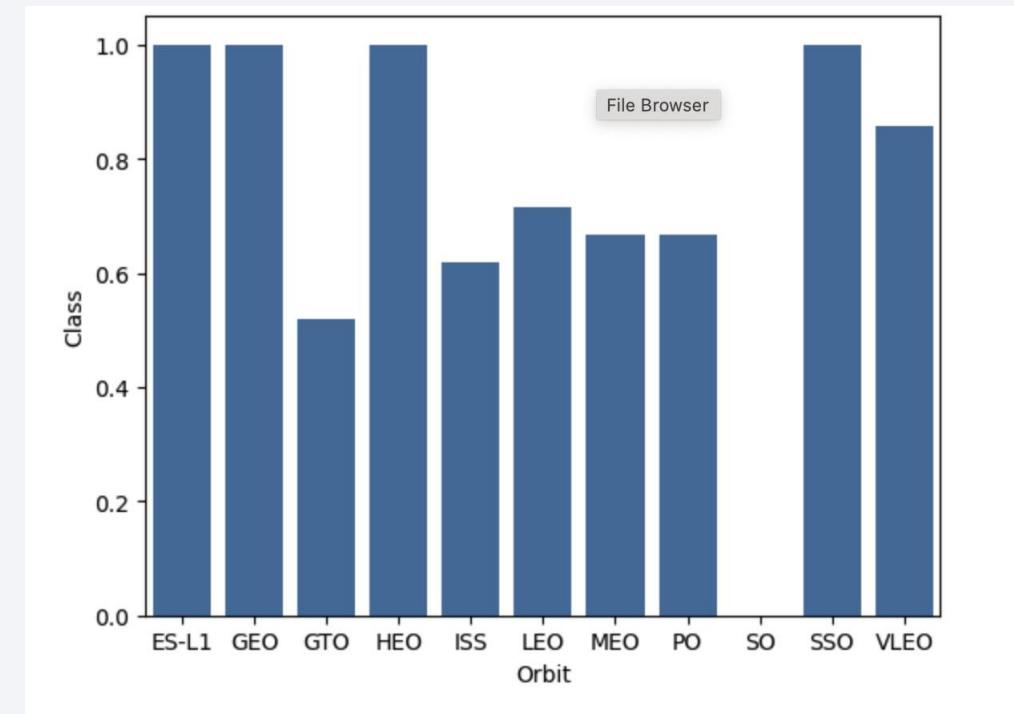
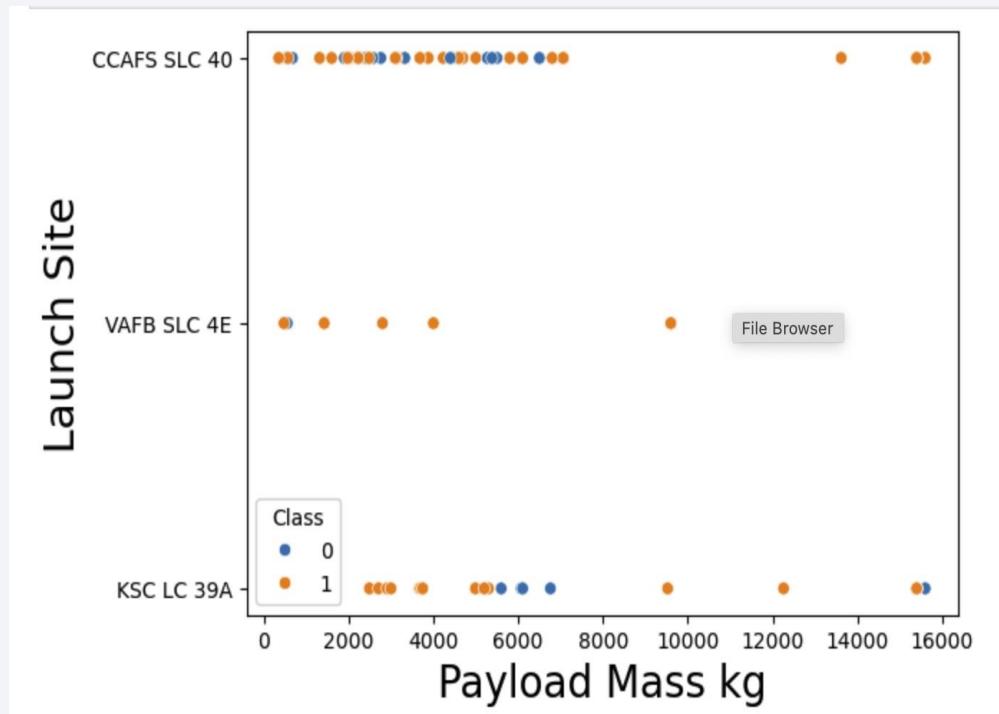
# EDA with Data Visualization

Charts to visualize relationship between Flight Number and Payload/Launch Site



# EDA with Data Visualization

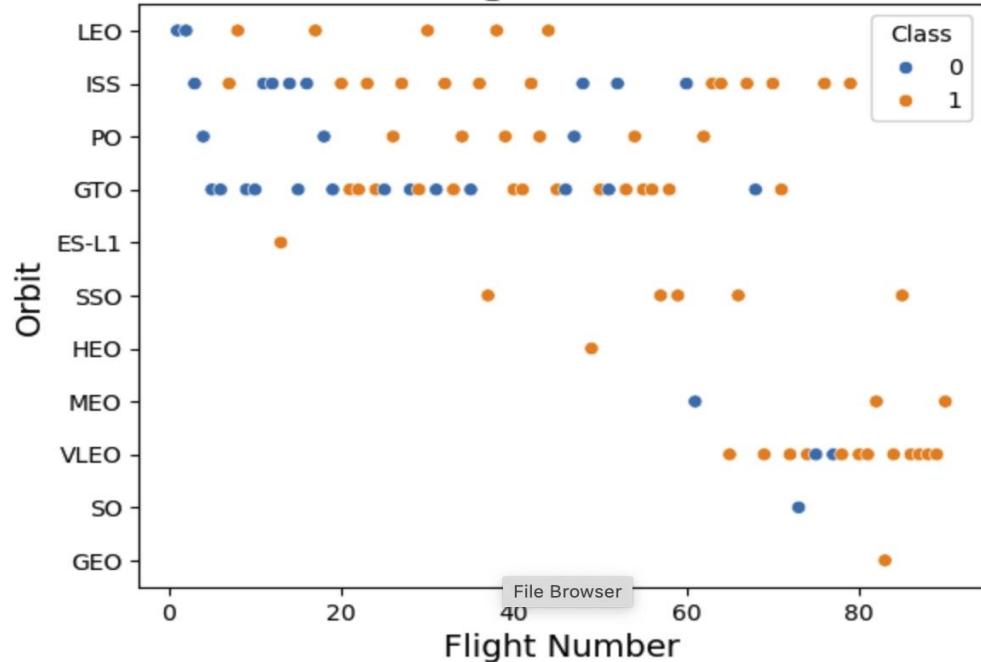
Charts to visualize the effects of each attributes



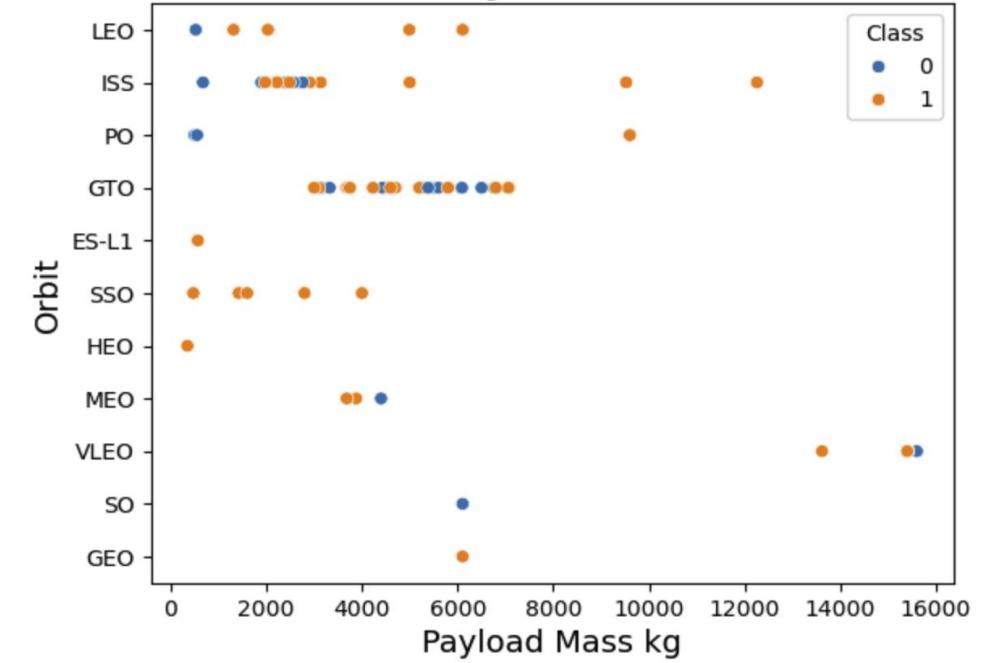
# EDA with Data Visualization

Charts to visualize the effects of each attributes

Relation between Flight Number and Orbit Type



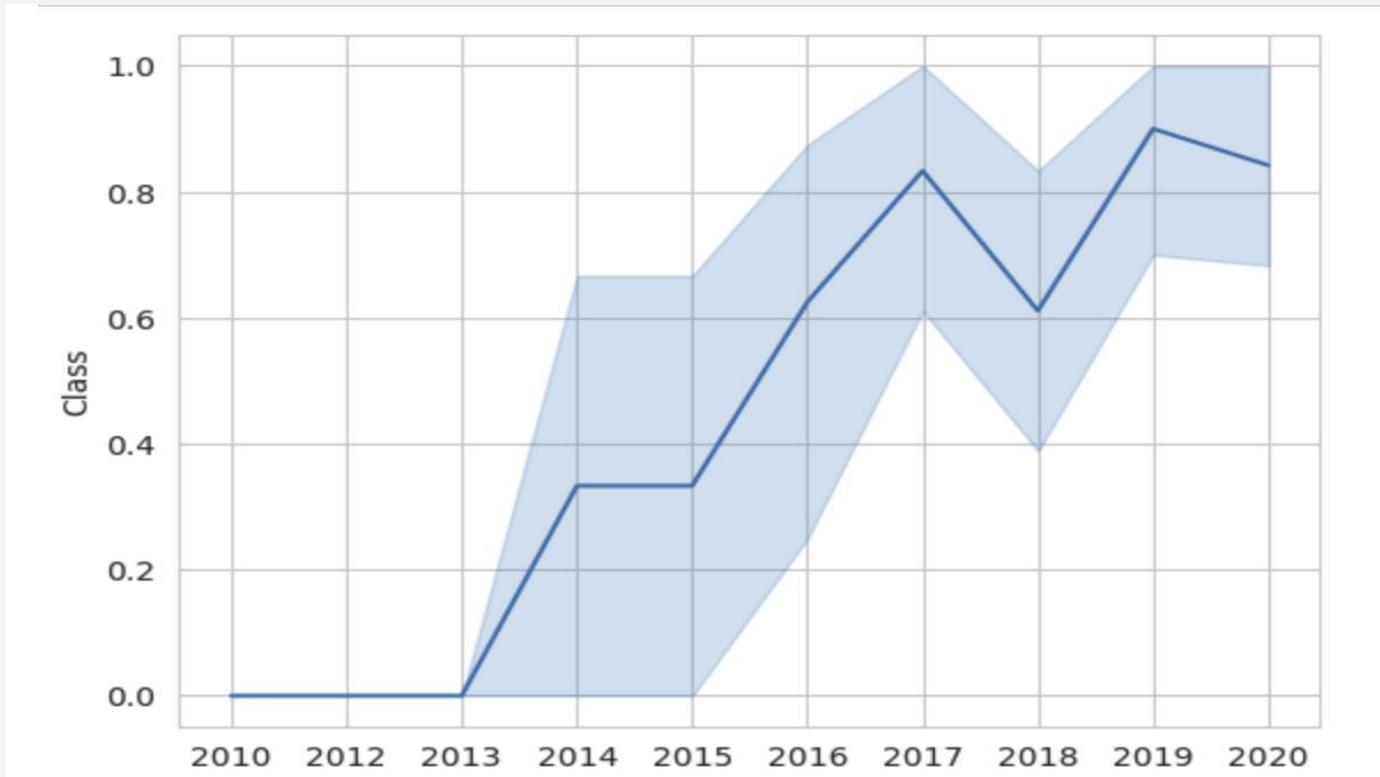
Relation between Payload Mass and Orbit Type



# EDA with Data Visualization

---

The progress of SpaceX between 2010 and 2020



<https://labs.cognitiveclass.ai/v2/tools/jupyterlite?ulid=ulid-bcb171190569d4ac32e7a0fba87ff5868cbfdd09>

# EDA with SQL

---

## SQL queries performed:

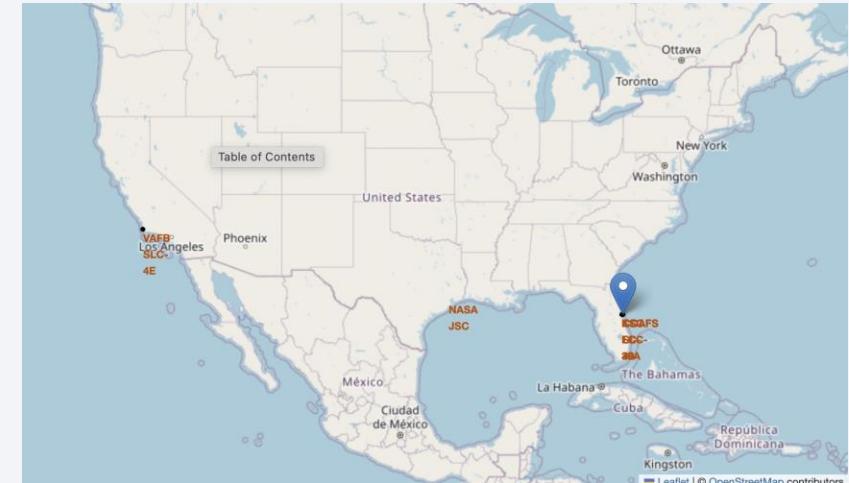
- Display the unique launch sites
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload greater than 400 but less than 6000
- List the total number of successful and failure mission outcomes
- List all the booster\_versions that have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship, booster version, launch site for 2015
- Rank the count of landing outcomes

[https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX\\_Data\\_Lab/jupyter-labs-eda-sql-coursera\\_sqlite.ipynb](https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX_Data_Lab/jupyter-labs-eda-sql-coursera_sqlite.ipynb)

# Build an Interactive Map with Folium

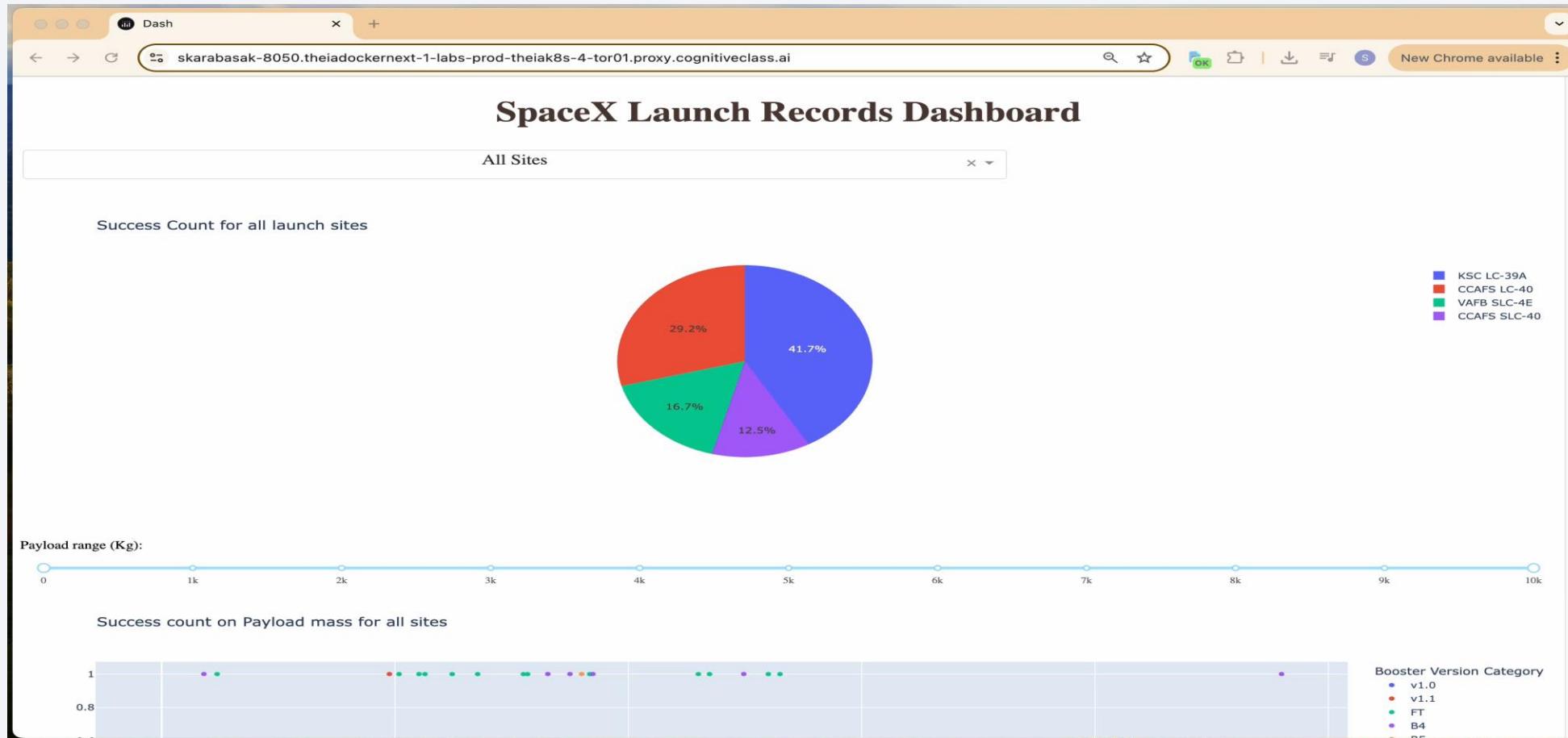
---

- Markers, circles, lines are created and added to a folium map
  - Circle is used to highlight the circle area with a text label on a specific coordinate
  - Color-labeled markers in marker cluster helps to easily identify which launch site have relatively high success rates
  - Mouse position gets coordinate for a mouse over a point on the map



[https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX\\_Data\\_Lab/lab\\_jupyter\\_launch\\_site\\_location.ipynb](https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX_Data_Lab/lab_jupyter_launch_site_location.ipynb)

# Build a Dashboard with Plotly Dash



# Predictive Analysis (Classification)

---

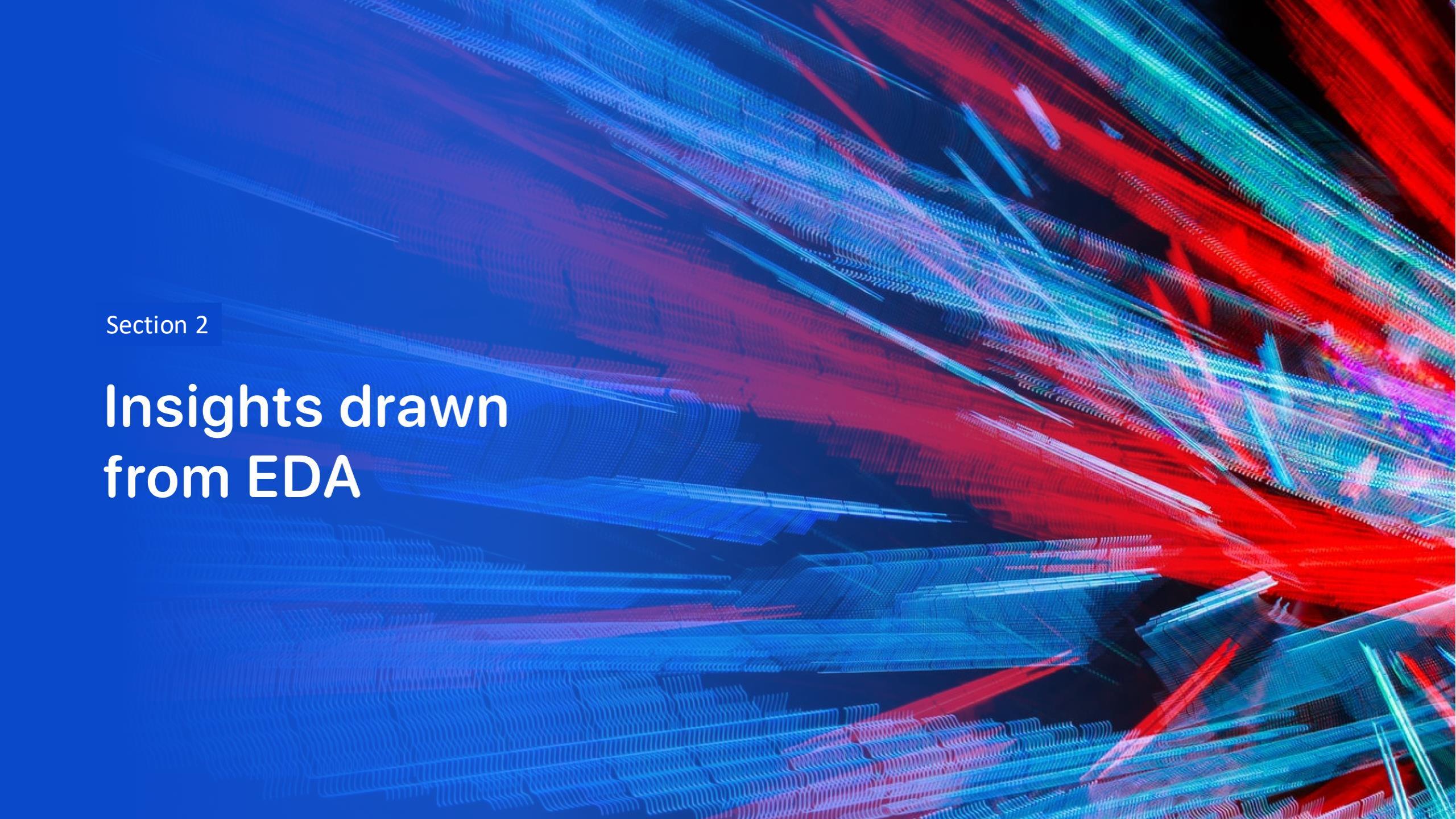
- Split the data into training testing data
- Train different classification models.
- Optimize the Hyperparameter grid search
- Find the the method that performs best using test data

[https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX\\_Data\\_Lab/SpaceX\\_Machine%20Learning%20Prediction\\_Part\\_5.ipynb](https://github.com/simge-git/IBM-SpaceX-Capstone-Project/blob/main/SpaceX_Data_Lab/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)

# Results

---

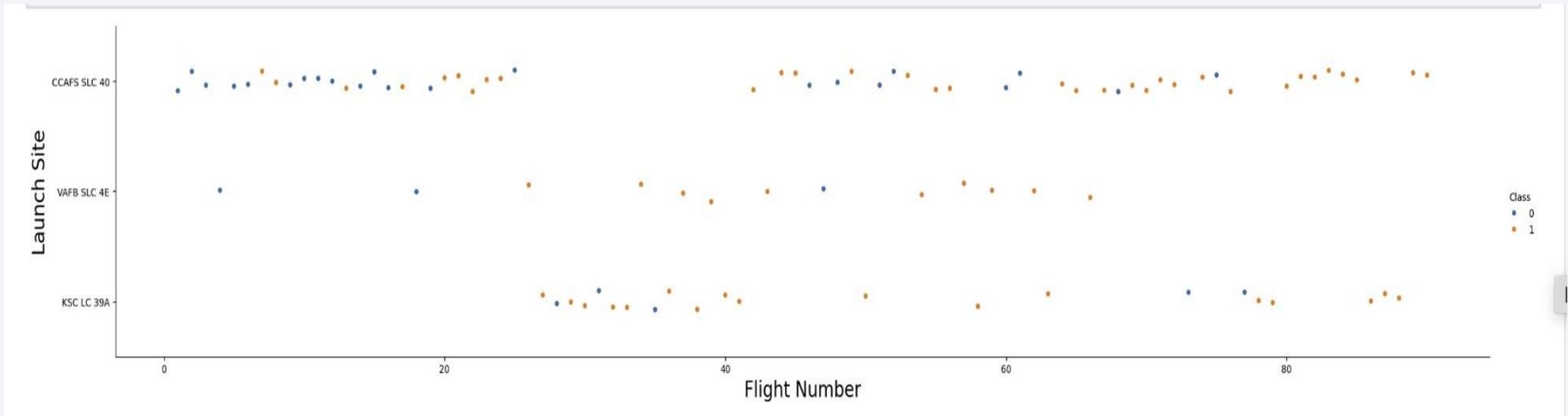
- Exploratory data analysis results:
  - Increase in flight number has positive effect on the success
  - There is no rockets launched for heavy payload mass (greater than 10000) at VAFB-SLC
  - ES-L1, GEO, HEO and SSO orbits have the highest success rates
  - Success rate kept increasing between 2013 and 2020
- The SVM, KNN and Logistic Regression models have best accuracy rate

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a 3D wireframe or a network of data points. The overall effect is futuristic and dynamic, suggesting concepts like data flow, digital communication, or complex systems.

Section 2

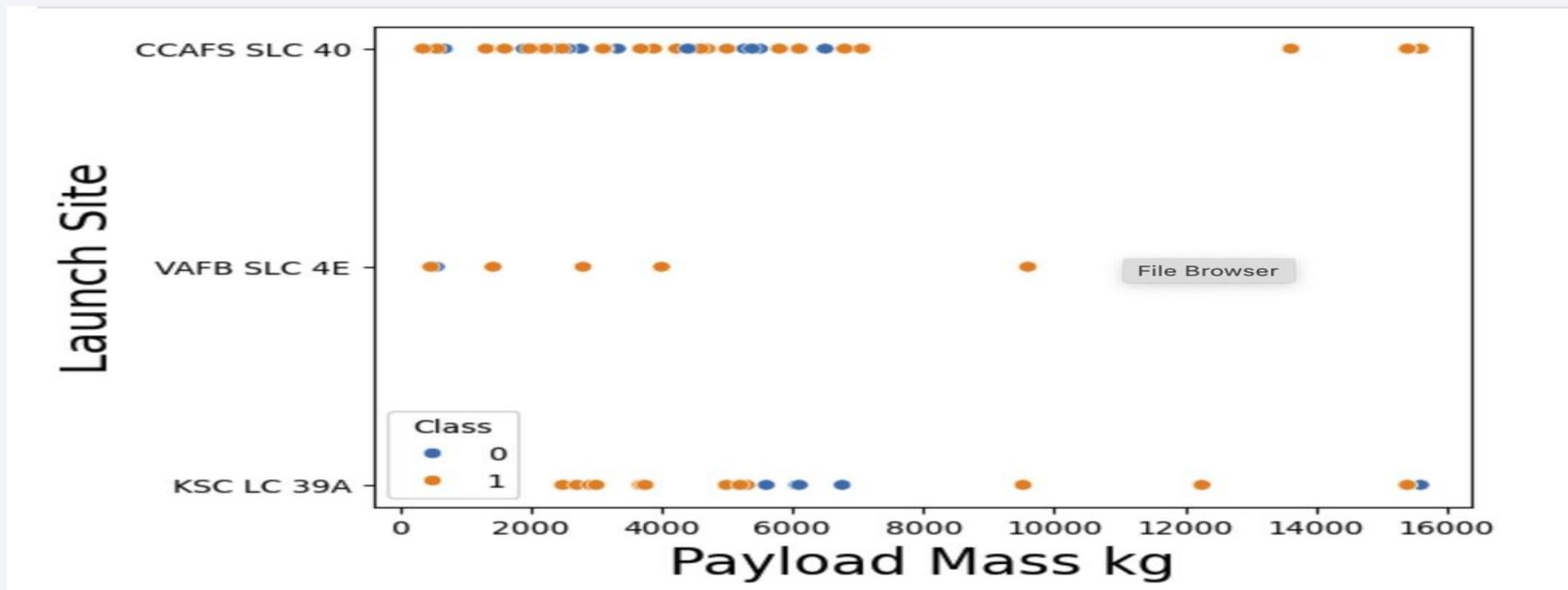
## Insights drawn from EDA

# Flight Number vs. Launch Site



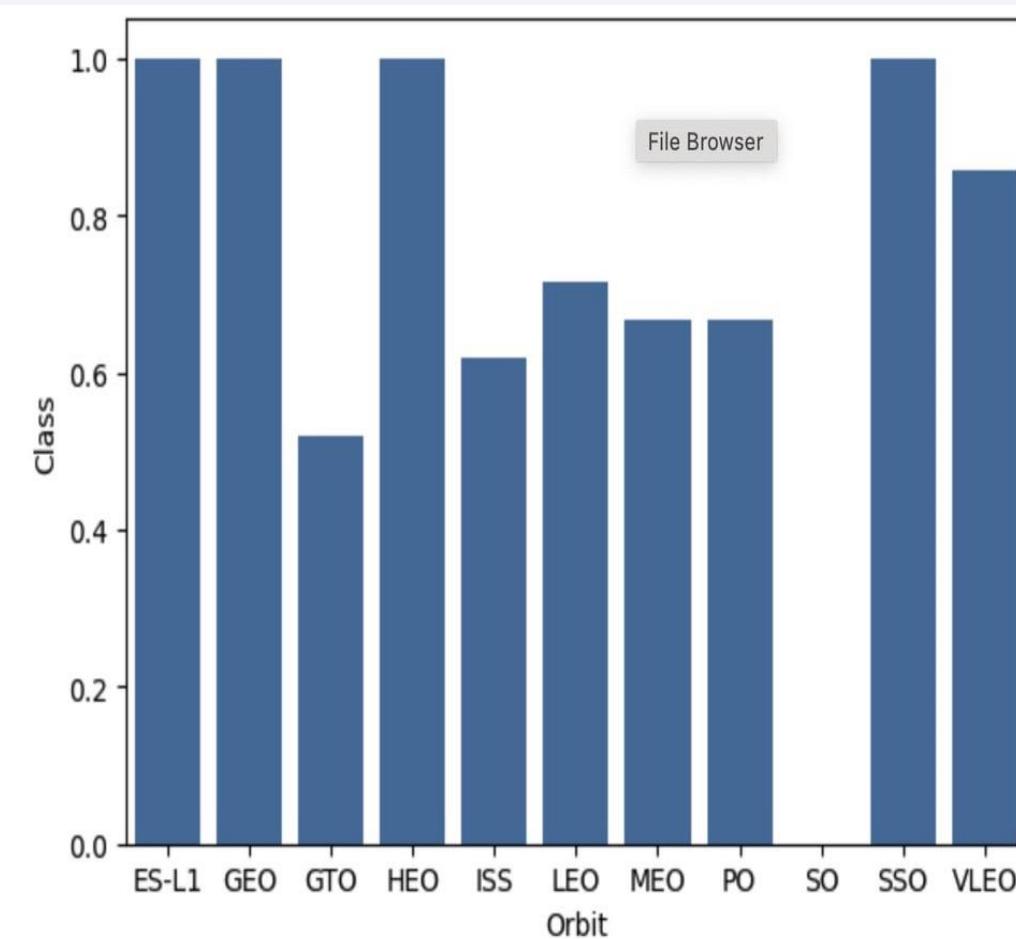
- When the number of flight increases, the success rate also increases

# Payload vs. Launch Site



There is no rockets launched for heavy payload mass (greater than 10000) at VAFB-SLC

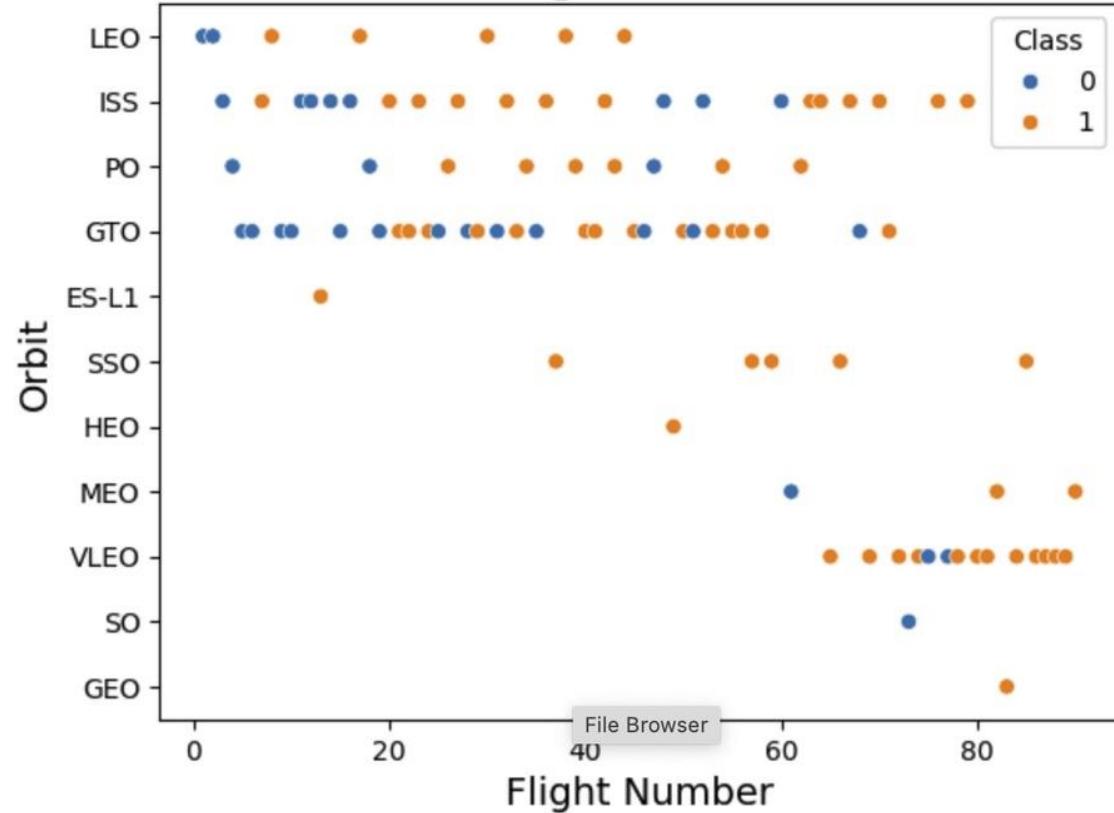
# Success Rate vs. Orbit Type



- ES-L1, GEO, HEO and SSO orbits have the highest success rates

# Flight Number vs. Orbit Type

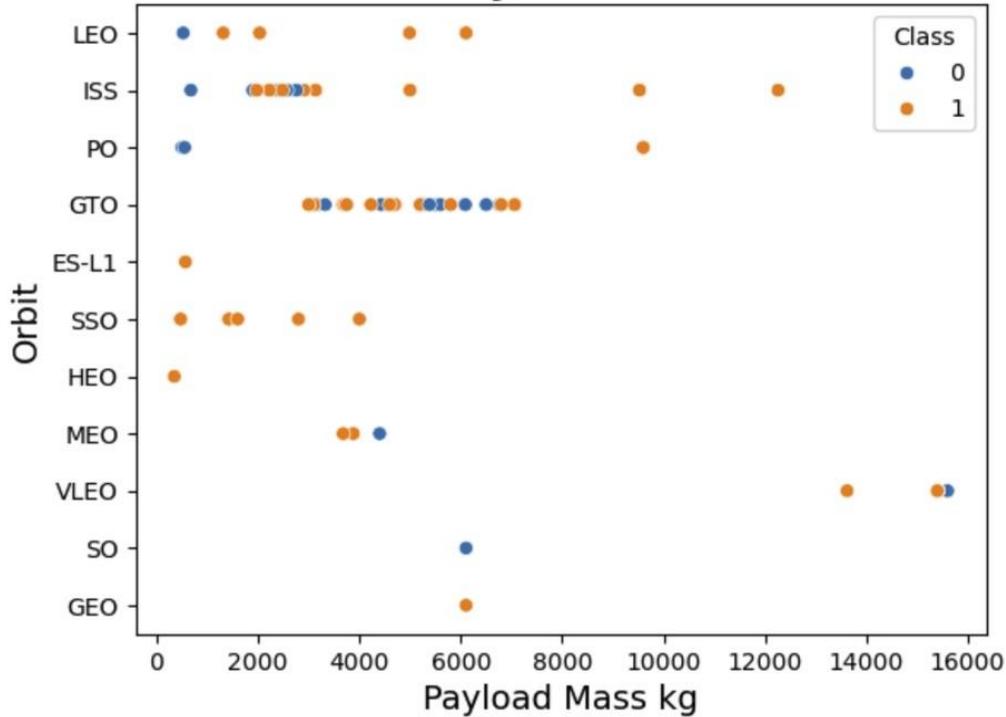
Relation between Flight Number and Orbit Type



- In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success

# Payload vs. Orbit Type

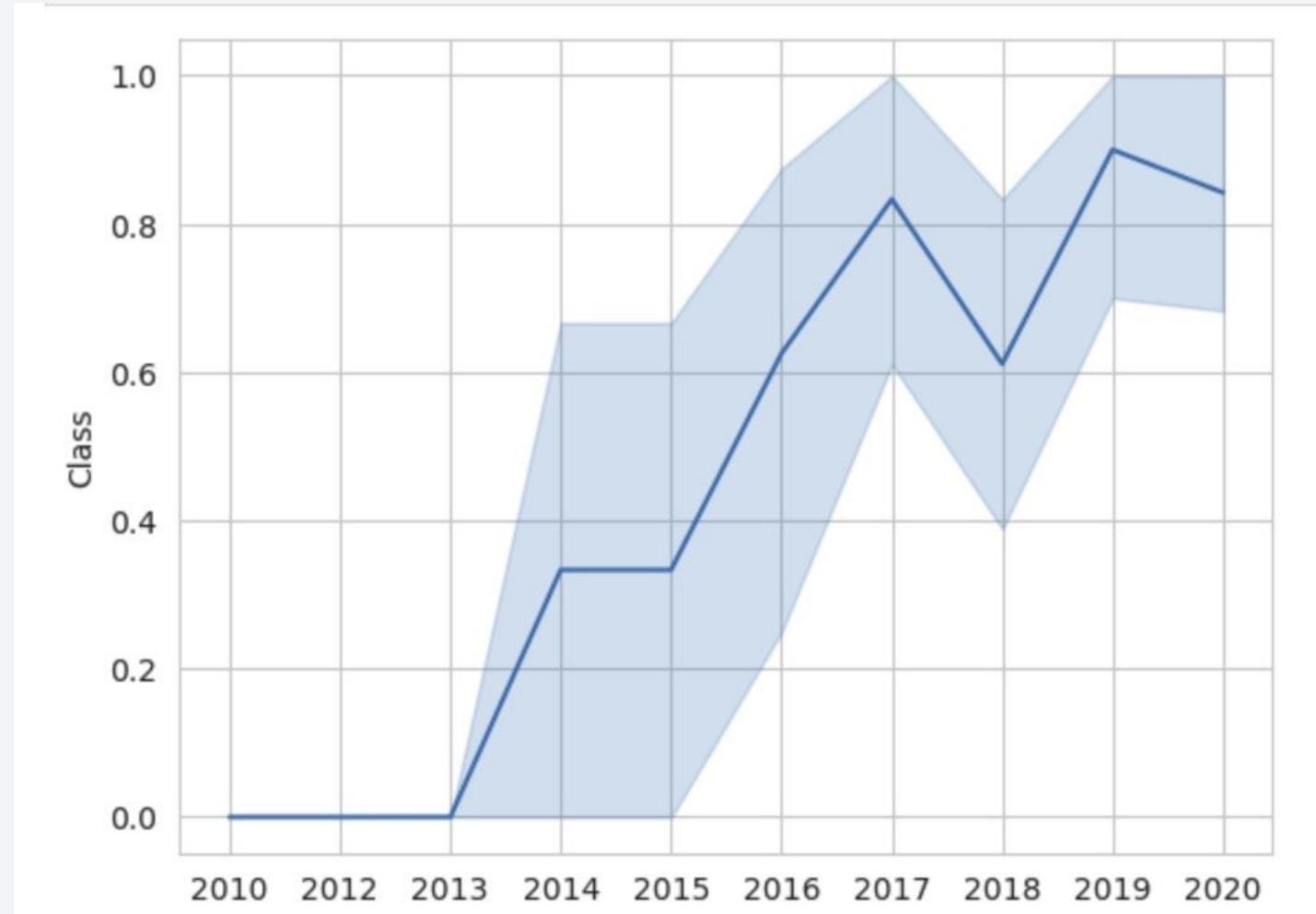
Relation between Payload Mass and Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

# Launch Success Yearly Trend

---



- Launch success rate kept increasing between 2013 and 2020

# All Launch Site Names

---

```
: %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'

---

```
: %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
```

# Total Payload Mass

---

```
: %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where "Booster_Version" like "F9 v1.1%"  
* sqlite:///my_data1.db  
Done.  
: AVG(PAYLOAD_MASS__KG_)  
: _____  
: 2534.666666666665
```

# Average Payload Mass by 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

---

```
: %sql select min("Date") from SPACEXTBL where "Landing_Outcome" = "Success (ground pad)"  
:  
* sqlite:///my_data1.db  
Done.  
:  
min("Date")  
-----  
2015-12-22
```

Name: jupyter-labs-eda-sql-  
coursera\_sqlite.ipynb  
Size: 35.9 KB  
Path: DS0321EN/labs/module\_2/SQLLite  
Created: 7/4/25, 9:48 PM

## Successful Drone Ship Landing with Payload between 4000 and 6000

---

```
: %sql Select "Booster_Version" from SPACEXTBL where "Landing_Outcome"="Success (drone ship)" and "PAYLOAD_MASS_KG_" between 4000 and 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

---

```
: %sql select count (Mission_Outcome) as missionoutcome from SPACEXTBL Group by Mission_Outcome  
* sqlite:///my_data1.db  
Done.  
: missionoutcome  
-----  
1  
98  
1  
1
```

# Boosters Carried Maximum Payload

---

```
%sql select Booster_version as BoosterVersion, PAYLOAD_MASS__KG_ as payload from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
* sqlite:///my_data1.db
Done.

BoosterVersion    payload
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
```

# 2015 Launch Records

```
%sql Select CASE SUBSTR("Date", 6, 2) when "01" then "January" when "02" then "February" when "03" then "MArch" when "04" then "April" when "05" then "May" when "06" then "June" when "07" then "July" when "08" then "August" when "09" then "September" when "10" then "October" when "11" then "November" when "12" then "December" end as month, "Landing_Outcome" = "Failure (drone ship)" as Landing_Outcome, Booster_Version, Launch_Site  
* sqlite:///my_data1.db  
Done.  


| month    | "Landing_Outcome" | Booster_Version | Launch_Site |
|----------|-------------------|-----------------|-------------|
| January  | 1                 | F9 v1.1 B1012   | CCAFS LC-40 |
| February | 0                 | F9 v1.1 B1013   | CCAFS LC-40 |
| MArch    | 0                 | F9 v1.1 B1014   | CCAFS LC-40 |
| April    | 1                 | F9 v1.1 B1015   | CCAFS LC-40 |
| April    | 0                 | F9 v1.1 B1016   | CCAFS LC-40 |
| June     | 0                 | F9 v1.1 B1018   | CCAFS LC-40 |
| December | 0                 | F9 FT B1019     | CCAFS LC-40 |


```

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

---

```
: %sql select "Landing_Outcome", count("Landing_Outcome") from SPACEXTBL where "Date" between 20100604 and 20170320 group by "Landing_Outcome"  
* sqlite:///my_data1.db  
Done.  
: 

| Landing_Outcome        | count("Landing_Outcome") |
|------------------------|--------------------------|
| Success (drone ship)   | 12                       |
| No attempt             | 12                       |
| Success (ground pad)   | 8                        |
| Failure (drone ship)   | 5                        |
| Controlled (ocean)     | 4                        |
| Uncontrolled (ocean)   | 2                        |
| Precluded (drone ship) | 1                        |

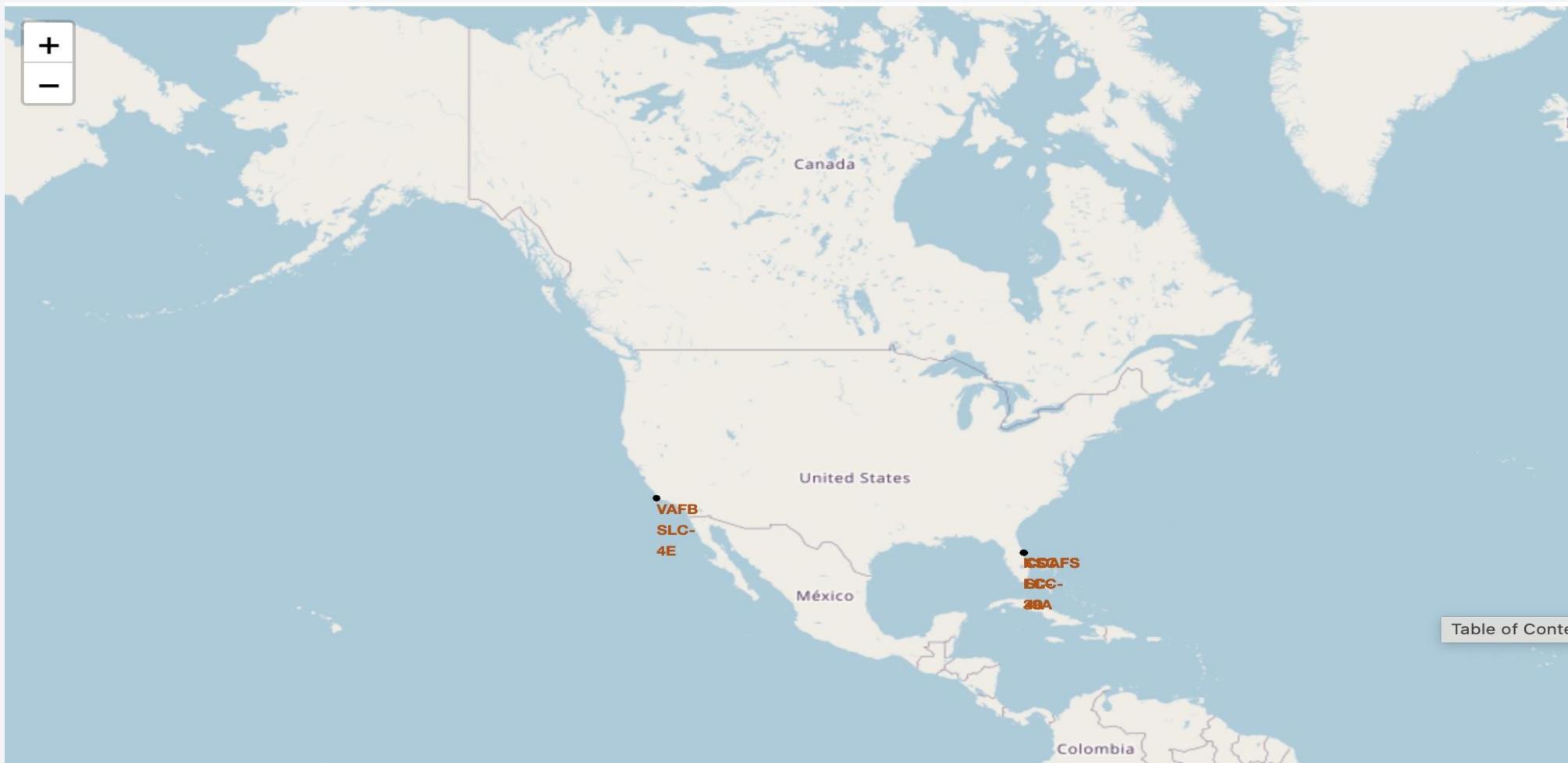

```

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth's horizon against a dark blue sky. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States appears. In the upper left quadrant, the green and yellow glow of the Aurora Borealis (Northern Lights) is visible.

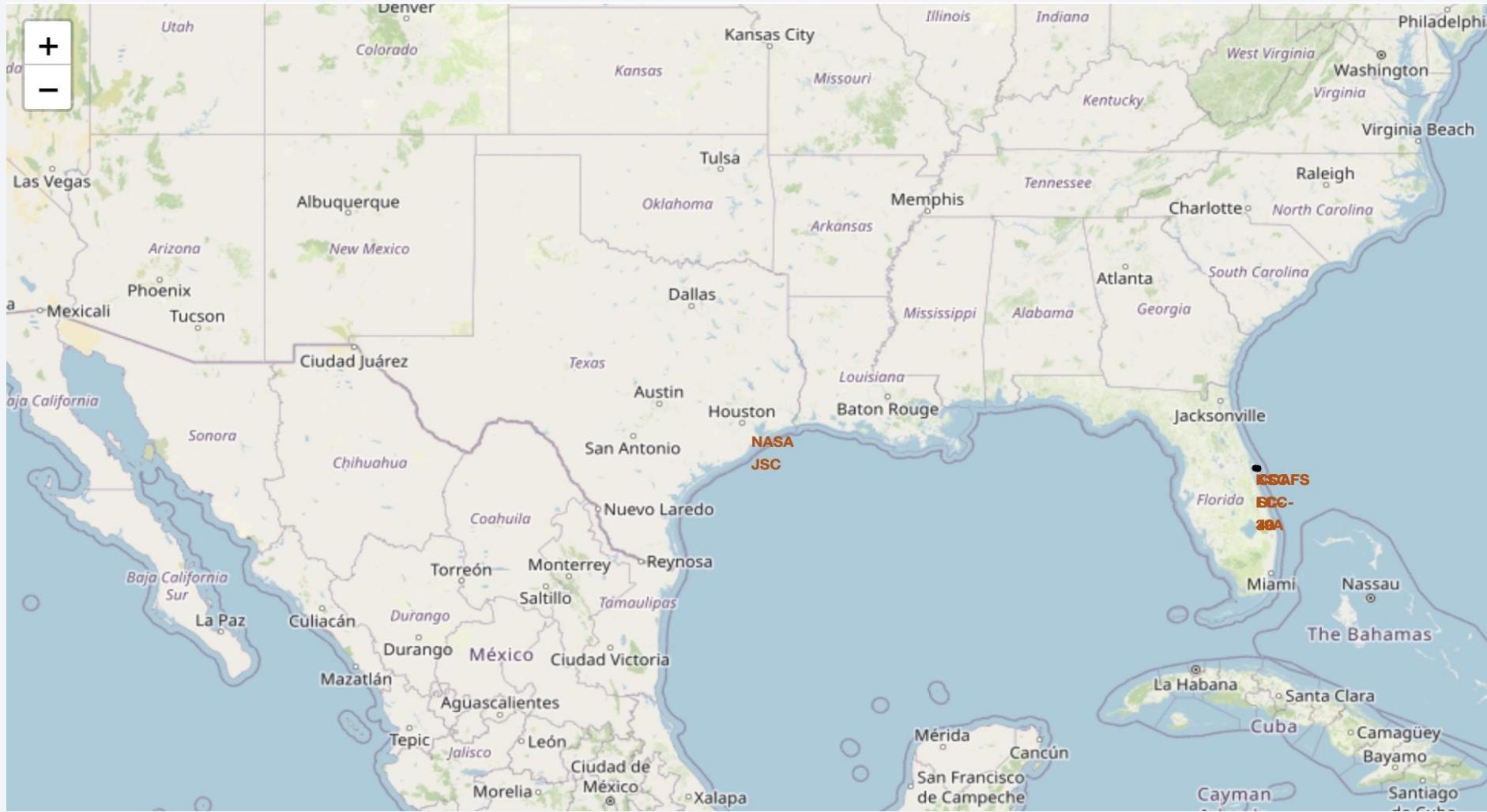
Section 3

# Launch Sites Proximities Analysis

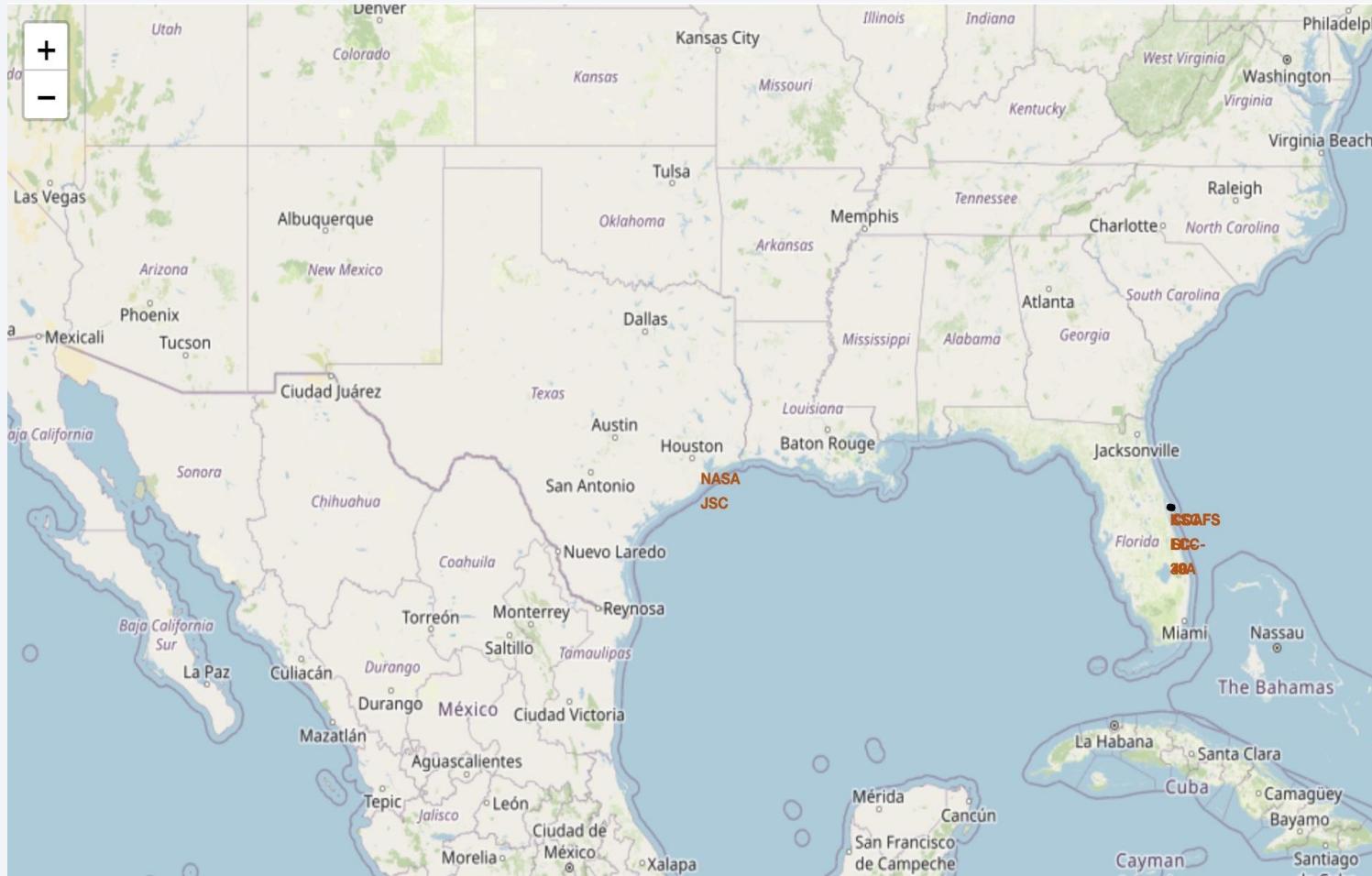
# All Launch Sites on a Global Map

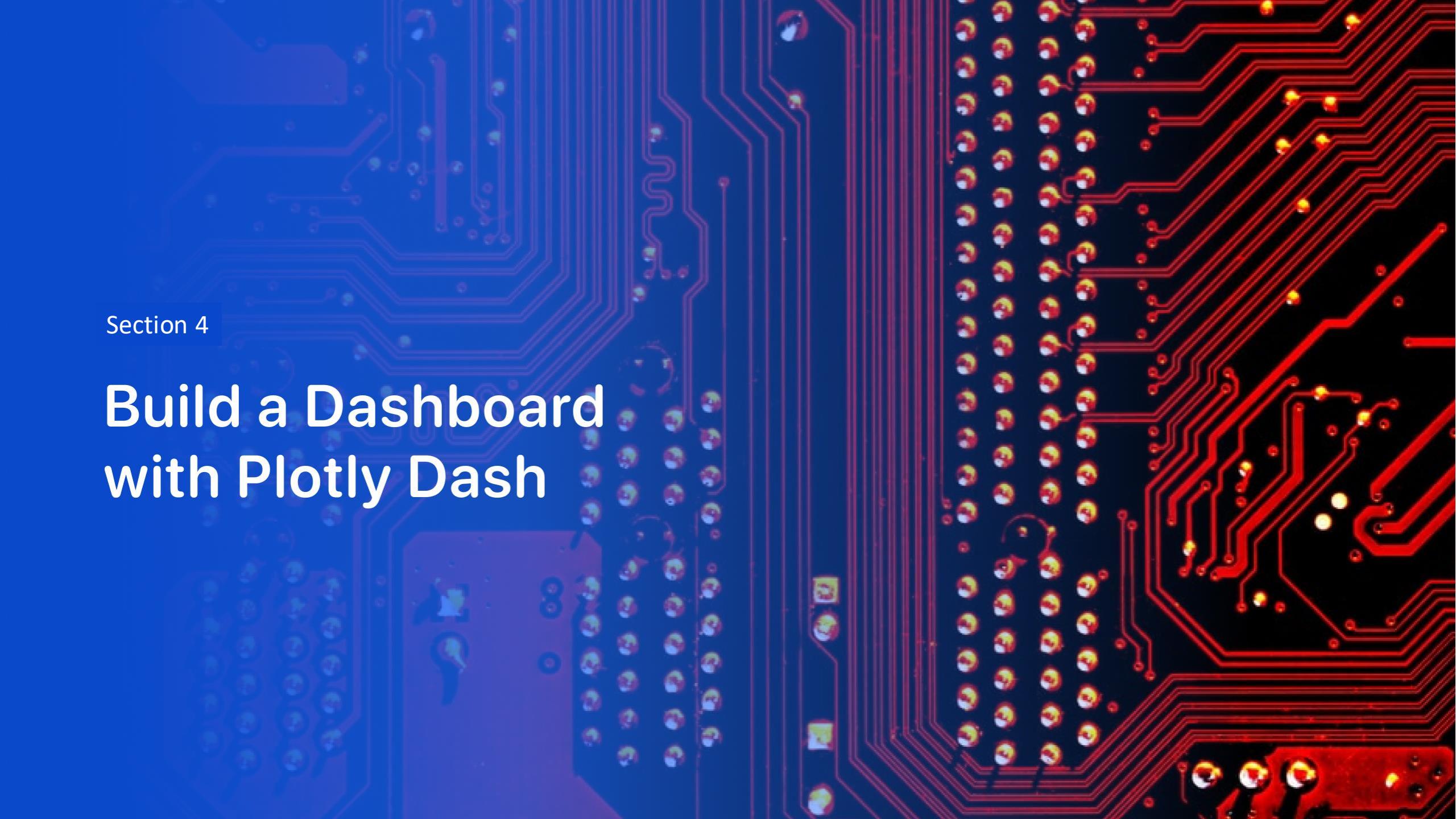


# Color-labeled Launch Outcomes on the Map



# Launch Site Proximities Map

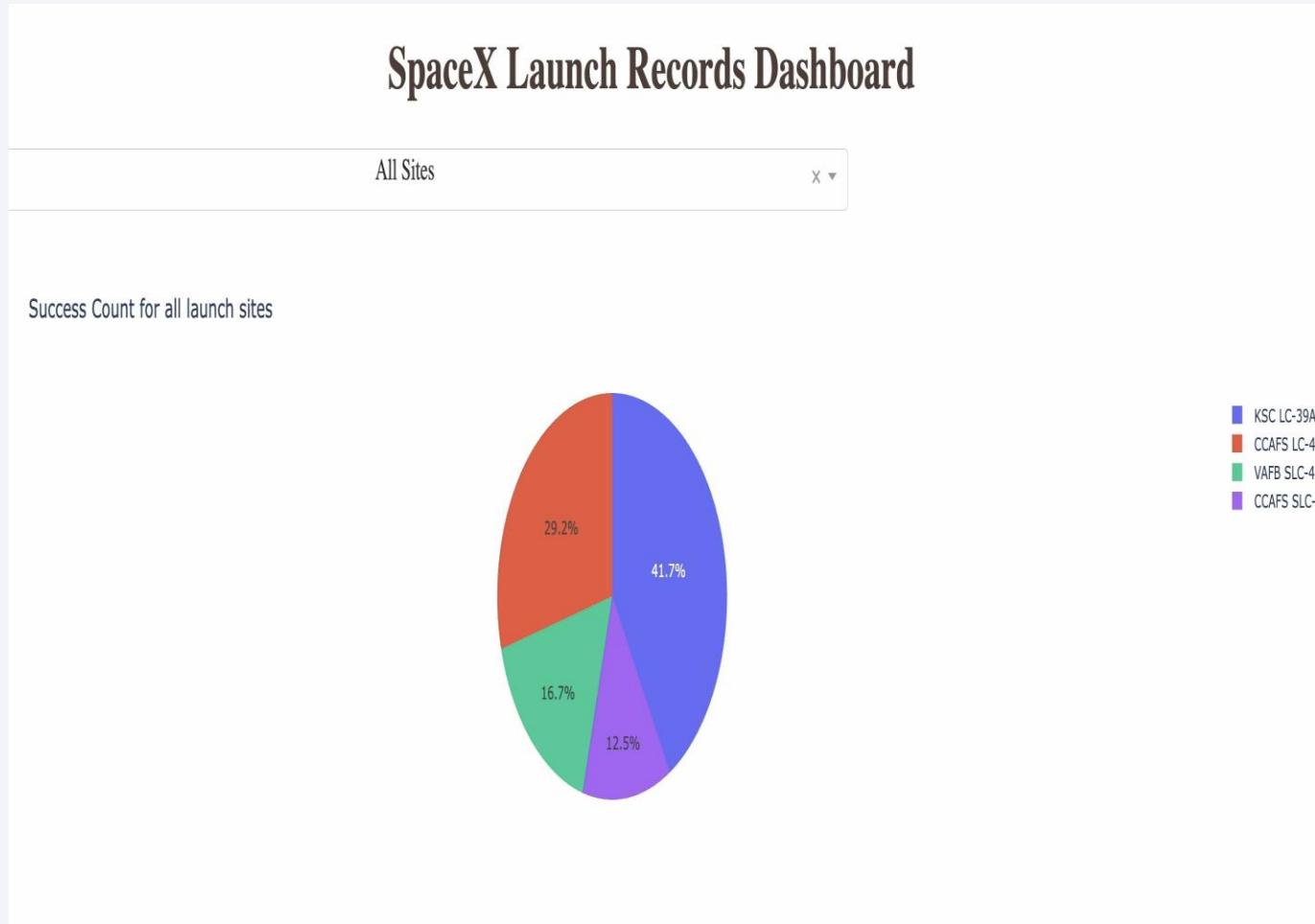


The background of the slide features a close-up photograph of a printed circuit board (PCB). The left side of the image has a blue color overlay, while the right side has a red color overlay. The PCB itself is dark grey or black, with numerous red and blue printed circuit lines (traces) connecting various components. Components visible include a large blue integrated circuit package at the top left, several smaller yellow and orange components, and a grid of surface-mount resistors on the left edge.

Section 4

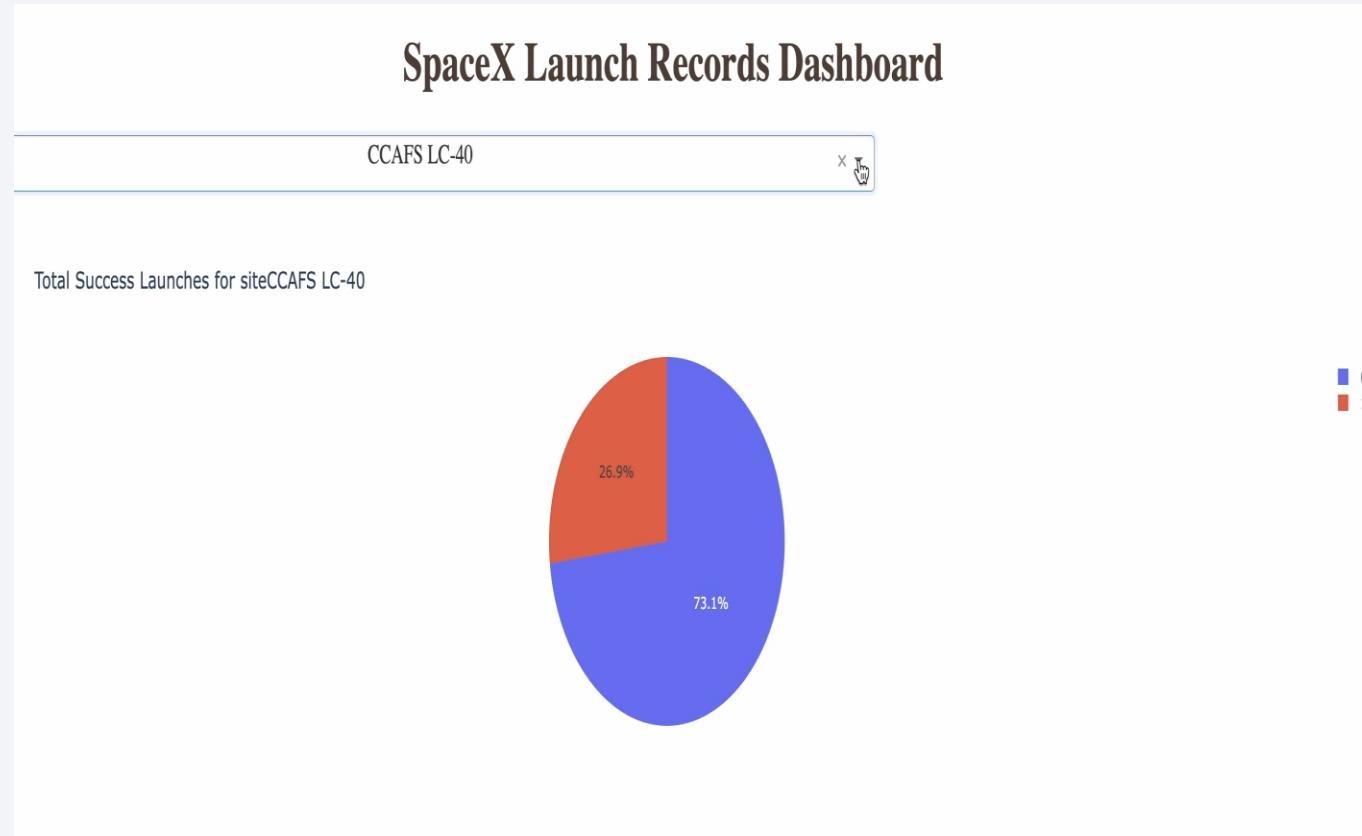
# Build a Dashboard with Plotly Dash

# Success Count for All Sites



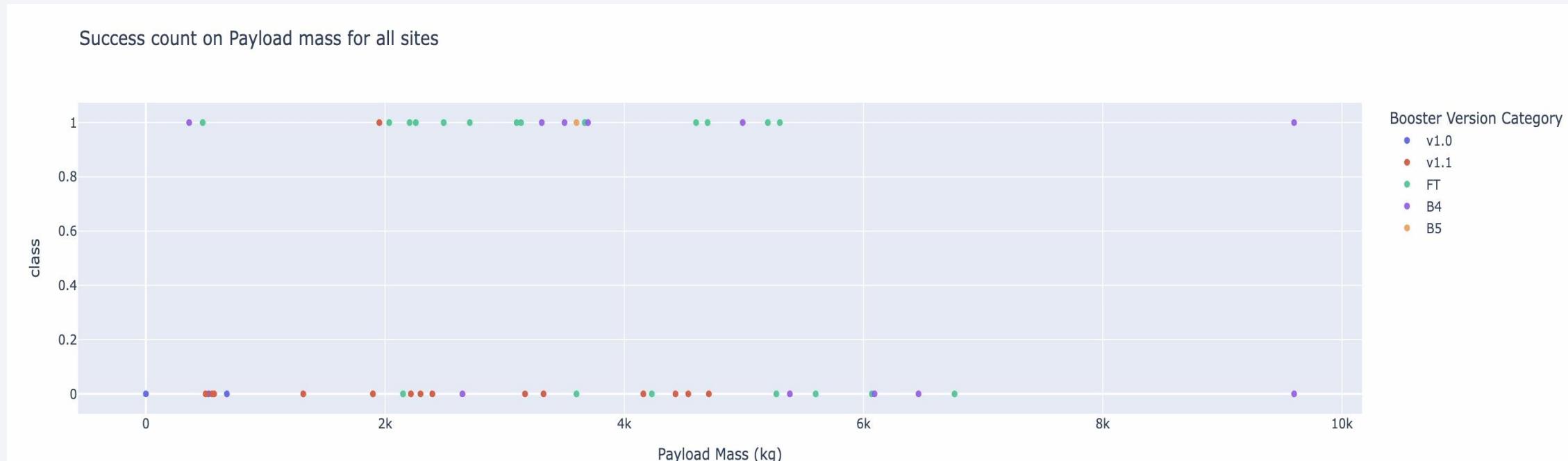
- KSC LC-39A has the most successful launches rate with 41.7%

# Launch Success Ratio for CCAFS LC-40



- The second most successful Launch Site  
CCAFS LC-40 has % success rate

# Payload vs. Launch Outcome for All Sites



The success rates for low-weighted payloads is higher than the heavy-weighted payloads

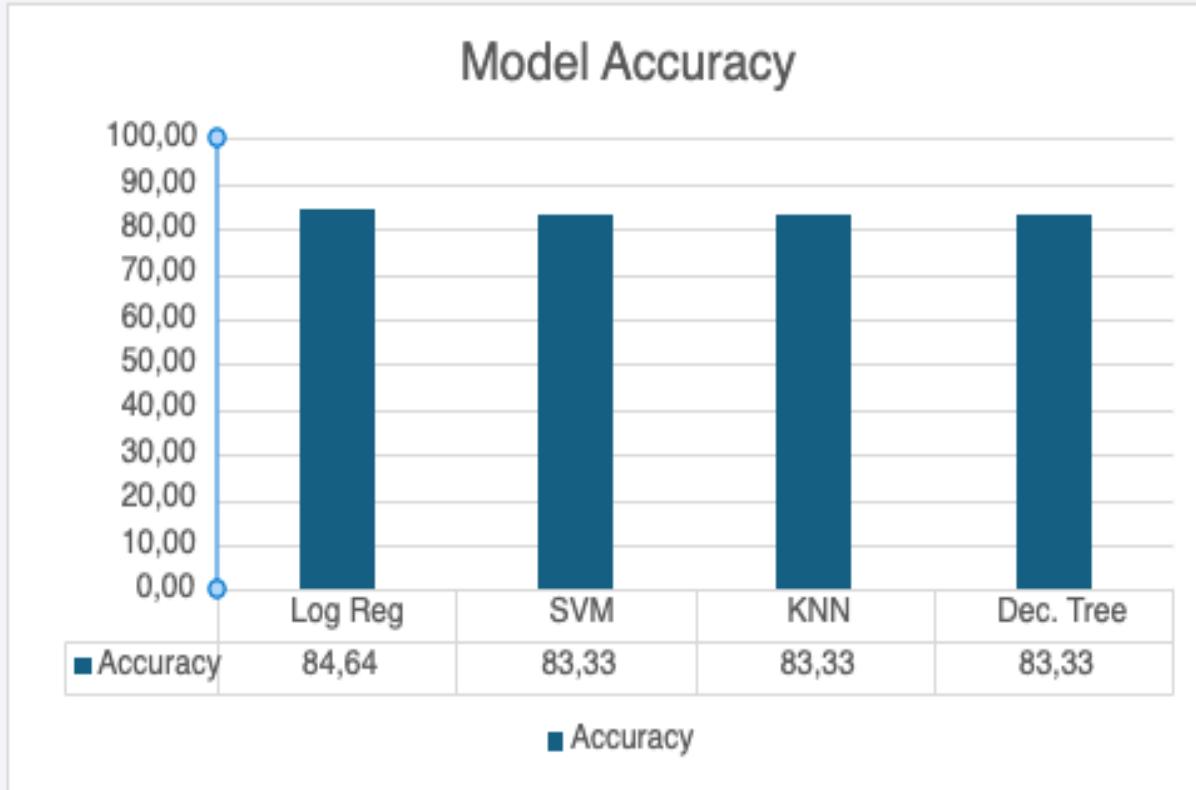
The background of the slide features a dynamic, abstract design. It consists of several thick, curved lines in shades of blue and yellow, creating a sense of motion and depth. The lines curve from the bottom left towards the top right, with some lines being more prominent than others. The overall effect is reminiscent of a tunnel or a high-speed journey through a digital space.

Section 5

# Predictive Analysis (Classification)

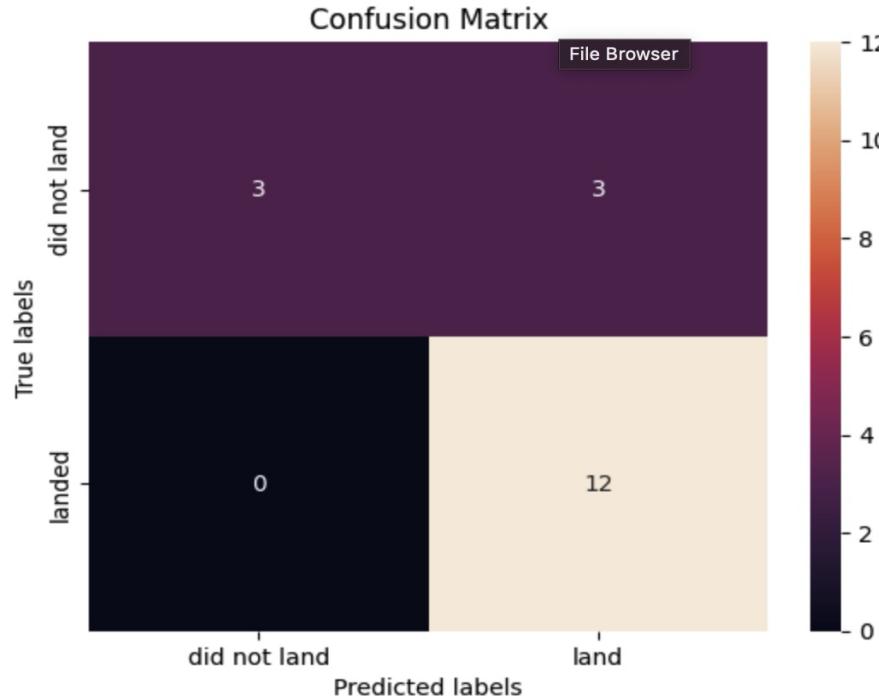
# Classification Accuracy

---



Logistic Regression model has the highest classification accuracy

# Confusion Matrix



Logistic Regression  
Confusion matrix

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.

Overview:

True Positive - 12 (True label is landed, Predicted label is also landed)

False Positive - 3 (True label is not landed, Predicted label is landed)

# Conclusions

---

- Logistic Regression model has better accuracy
- There are 12 True Positive and 3 False Positive

Thank you!

